A Role for Introspection in AI Research

Samuel Freed

November 2016

Thesis submitted in partial fulfilment for the degree of

Doctor of Philosophy in Informatics

School of Engineering and Informatics

University of Sussex

Abstract

The main thesis is that **Introspection is recommended for the development of anthropic AI**.

Human-like AI, distinct from rational AI, would suit robots for care for the elderly and for other tasks that require interaction with naïve humans. “Anthropic AI” is a sub-type of human-like AI, aiming for the pre-cultured, universal intelligence that is available to healthy humans regardless of time and civilisation. This is contrasted with western, modern, well-trained and adult intelligence that is often the focus of AI. Anthropic AI would pick up local cultures and habits, ignoring optimality. Introspection is recommended for the AI *developer*, as a source of ideas for designing an artificial mind, in the context of technology rather than science. Existing notions of introspection are analysed, and the aspiration for “clean” or “good” introspection is exposed as a mirage. Nonetheless, introspection is shown to be a legitimate source of ideas for AI using considerations of the contexts of discovery vs. justification. Moreover, introspection is shown to be a positively plausible basis for ideas for AI since if a teacher uses introspection to extract mental skills from themselves to transmit them to a student, an AI developer can also use introspection to uncover the human skills that they want to transfer to a computer. Methods and pitfalls of this approach are detailed, including the common error of polluting one's introspection with highly-educated notions such as mathematical methods.

Examples are coded and run, showing promising learning behaviour. This is interpreted as a compromise between Classic AI and Dreyfus's tradition. So far AI practitioners have largely ignored the subjective, while the Phenomenologists have not written code – this thesis bridges that gap. One of the examples is shown to have Gadamerian characteristics, as recommended by (Winograd & Flores, 1986). This serves also as a response to Dreyfus's more recent publications critiquing AI (Dreyfus, 2007, 2012).

To

my brother, who dedicated his Ph.D. thesis to our parents, who dedicated their Ph.D. theses to their parents.

Acknowledgements

From the time before my arrival in Sussex, I wish to thank Oron Shagrir for his ongoing support, David Harel for pointing out that I have to go abroad to study AI at doctoral level, and Michael Wheeler for his unremitting efforts at finding me a place to do my somewhat unusual Ph.D. work and for recommending Sussex.

Within Sussex, I first wish to thank my main supervisor, Ron Chrisley, for endless hours of detailed feedback, and for even in the hardest of times never giving up on this project. I am immeasurably a better scholar now thanks to his patience. I thank my secondary supervisor Chris Thornton for many lunches full of humorous guidance and key advice. My Ph.D. committee member, Steve Torrance, has also been an important source of feedback.

Amongst the other members of the department I thank especially David Schwartzman and Blay Whitby for their encouragement, good humour, friendship and appreciation of my coffee preferences over the years. I also thank Jens Streck and Esin Yavuz, my office-mates, and my very international group of friends in and around Sussex, who all suffered my philosophical rants and changing moods with great patience.

I also wish to thank Peter Cheng and David Berry (from the digital humanities) for encouragement and clear guidance in moments of despair.

Back in Israel I wish to thank Eli Sandler, Gilad Landau and Josh Weinstein for long and insightful discussions over many years that have significantly advanced my thinking on the subjects discussed here. Last but not least my family has given me, *inter alia*, much encouragement and decades-long toleration of my obsession with the ideas in this thesis.

Table of contents

1 Introduction, context and methodology....1
1.1 Motivations....2
1.2 Outlines....5
1.3 The specific field of this thesis....12
1.4 Notions of truth....15
1.5 Science vs technology and human-like vs rational....21
2 Literature review....23
2.1 The Cognition-vs-Phenomenology debate....24
2.2 Simon....28
2.3 Dreyfus....31
2.4 Winograd & Flores....38
2.5 Hermeneutics and Gadamer....43
2.6 Literature review: summary....47
3 Thesis outline and terms....49
3.1 Terms of this thesis: “is recommended for developing”....50
3.2 Terms of this thesis: “anthropic”....55
3.3 Terms of this thesis: “introspection”....76
4 Introspection may legitimately be used for AI....94
4.1 Introspection as “impossible”....96
4.2 Introspection as “forbidden”....97
4.3 Introspection as “commonplace”....105
4.4 Introspection as “desirable”....112
4.5 Introspection as “unavoidable”....114
4.6 A hybrid position....115
4.7 Types of truth in introspection....117
4.8 Introspection may legitimately be used for AI: summary....121
5 Introspection is likely to be profitable....122
5.1 Conceptual arguments....123
5.2 An argument from education....124
5.3 Programming impossible without introspection....133
5.4 Introspection is likely to be profitable: summary....136
6 Details and how to use introspection for AI....138
6.1 Definitions and delineations....139
6.2 The process of introspection for AI....146
6.3 Comments on the process of introspection for AI....148
6.4 Project expectations....158
6.5 Testing and evaluation....159
7 Examples of introspection being used for AI design....161

7.1 Fuzzy logic....164
7.2 Case based reasoning (CBR)....166
7.3 AIF0....168
7.4 AIF1....176
7.5 AIF2....177
7.6 Consequences of the examples....190
7.7 Examples of introspection being used for AI design: summary....193
8 Conclusion & possible consequences....194
8.1 Conclusion....195
8.2 Future technical work....198
8.3 Possible consequences for cognitive science....200
8.4 “Underpinning” models in philosophy....202
8.5 Open Questions....206
9 Appendix - US Patent No. 8,660,670....209
10 Bibliography....225

List of Figures

Illustration Index

Illustration 1.1: Locating different AI fields in key distinctions.....21
Illustration 3.1: Roles in AI development.....54
Illustration 6.1: Process of introspection for AI.....146
Illustration 7.1: Fuzzy Logic (Source: Wikimedia).....164
Illustration 7.2: Statistics on the Learning of AIF0 on the ABCD -> 1234 problem.....172
Illustration 7.3: AIF2 Data Types.....181

1 Introduction, context and methodology

Table of Contents

1 Introduction, context and methodology....1
1.1 Motivations....2
1.2 Outlines....5
1.2.1 Scope....5
1.2.2 Argument outline....6
1.2.3 Main antagonists....8
1.2.4 Document structure....10
1.2.5 Novelty....10
1.3 The specific field of this thesis....12
1.3.1 Philosophy of AI....12
1.3.2 Philosophy of technology....13
1.4 Notions of truth....15
1.4.1 The idea of a single truth....16
1.4.2 Perspectivism....16
1.4.3 Perspectives, realities, agendas, Occam....18
1.4.4 In what sense is this thesis true?....20
1.4.5 Notions of truth: summary....20
1.5 Science vs technology and human-like vs rational....21

My thesis is that "**Introspection is recommended for developing anthropic AI**". The term "anthropic" will only be defined in detail in section 3.2. Preliminarily it is the part of *human-like* AI that approximates the un-enculturated part of the mind – The part that allows culture and learning to arise. This is *opposed* to the "western, modern, well-trained and adult" mind.

This chapter opens with discussion of the motivations for the research, and gives outlines of the document: How it fits in a larger project, an overview of the argument, a list of antagonists, an overview of the document's structure, and a list of points of novelty.

To give some background to the entire argument, a discussion is due of the field of philosophy of AI specifically and of technology in general (section 1.3), and out of that come some concerns about various possible meanings of "truth" (section 1.4) (section 1.4.4 will discuss the type of truth I ascribe to the thesis itself). This is followed by a

delineation of the area within the field of AI that is being addressed (section 1.5).

In the next chapter there will be a literature review, and in chapter 3 the main terms of the thesis are discussed, such as introspection (section 3.3), recommendation (section 3.1.1), something being a basis for a technology (section 3.1.2), and AI development as such (section 3.1.3). Anthropic AI, the aim, is introduced in section 3.2.

1.1 Motivations

Much is being made in popular media about the recent successes of AI: computers are now the world champions in chess, go, and "jeopardy". Handwriting is being recognized, computer translation has improved, and many more aspects of AI that were confined to research labs are now in many pockets. However, most if not all of these recent achievements were done more by adding hardware and "brute force" to existing AI concepts than by entirely new concepts.

This work is motivated by this dearth of ground-breaking *new* ideas in AI. The need for this project can be illustrated by two frustrations and two opportunities:

The **first frustration** is, as mentioned above, that AI suffers a dearth of ground-breaking new ideas. The opinion that AI has been "*brain dead*" since at least the 1970s is supported by mainstream researchers such as Marvin Minsky (McHugh & Minsky, 2003).

An illustrative parallel to my concern with AI could be a "30,000 feet view" of the development of *fundamentally new ideas* in wheeled transport. Though no-one would doubt that wheeled transport has made great strides in the 20th century, trains were invented in the early 19th century, and became widely deployed towards the end of that century. Similarly, the bicycle was developed during most of the 19th century from an early concept to models that we would recognise as quite modern by the end of that period. Both the car and the motorbike are creatures of the late 19th century. Arguably the next fundamental *conceptual* innovation in wheeled transport was the "Segway", introduced in 2001. So surprisingly, we had no *conceptual* innovations in wheeled transport in the entire period of the 20th century – only incremental progress. This is not to belittle the efforts of automotive engineers in the 20th century – just to point out that there were few if any new *fundamental ideas* that caught on. The first frustration here is

that from this same 30,000-feet view we have had very little progress in AI in recent decades.

The fundamental ideas of symbolic AI were already in place in the 1950s (McCorduck, 2004). The basics of neural nets are found in (McCulloch & Pitts, 1943), and were developed as a mainstay of AI by the PDP group in the 1970s (Nilsson, 2010, p. 339). The ideas of statistical AI can be seen as an extension of symbolic and mathematical AI. Conceptual innovations are few, far between, and not that successful, with the possible exception of the work of Brooks on minimally-intelligent systems, but these make no pretence to intelligence beyond that of an insect.

The **second frustration** (or perhaps a symptom of the first frustration) is that we have been led by popular culture to expect some sort of sentience, or at least human-like conduct from computers. This has completely failed to materialize. A case in point is “Star Trek (The Next Generation)”, a TV series released in 1987 which included the character “Data”, a robot that though not completely human-like, moved in human society, and took on human roles.

The above complaint about the absence of human-like AI is not only motivated by some romantic notions. Having human-like AI available to us would allow several applications of robotics that have a strong requirement for mutual understanding between untrained humans and robots. These applications require robots to think like humans in order to understand humans and be understood by humans, consider robotic care for the elderly (see section 3.2.2).

If the situation of AI is so disappointing, then maybe we ought to re-examine the boundaries of the discussion, so as to try to “think outside the box”. One of these boundaries is the way subjectivity and specifically introspection have been treated.

Understanding human thinking as such is part of the science of psychology, and is a long way in the future. But AI is a technological (and business) pursuit in the present, so we cannot wait for a full understanding of the human psyche to emerge. There are two sources of commentary on AI that make poignant points, yet have yet to lead to tangible results. Perhaps exploring these will provide an opportunity to get out of the impasse in the foundations of AI.

The first and better known of these **opportunities** is the tradition of critiquing AI from the standpoint of phenomenology (Dreyfus, 1979) and other backgrounds alien to the mainstream cognitive science (Winograd & Flores, 1986). Much of the field ignores these critiques (McCorduck, 2004, Chapter 9). Some of the literature review (chapter 2) is dedicated to understanding why the dialogue between this tradition and the mainstream of AI was so fruitless (Dreyfus, 2007).

A second opportunity is examining more closely the utterances of some central AI researchers when speaking more freely than seems possible in the peer-reviewed literature, e.g. (McCorduck, 2004; Turkle, 1984). When interviewed informally some researchers reveal severe tensions in their positions on subjective thinking in general and introspection in particular (see also sections 4.2 and 4.3). A revealing example of the paradoxes endemic to our thinking about thinking is given by Seymour Papert. This quote is key to understanding the context of this entire thesis:

> *We are to thinking as Victorians were to sex. We all know we have these horrible moments of confusion when we begin a new project, that nothing looks clear and everything looks awful, that we work our way out using all sorts of odd little rules of thumb, by going down blind alleys and coming back again, and so on, but since everyone else seems to be thinking logically, or at least they claim they do, then we figure we must be the only ones in the world with such murky thought processes. We disclaim them, and make believe that we think in logical, orderly ways, all the time knowing very well that we don't. And the worst offenders here are teachers, who present crisp, clean batches of knowledge to their students, and look as if they themselves had learned that knowledge in a crisp, clean way. It didn't happen that way, but the teachers don't admit it, and the students groan inwardly, feeling so hopelessly dumb.* (McCorduck, 2004, p. 339)

Papert describes a widespread tension in our culture, that we all pretend to be logical and sensible, while being acutely aware that this is not the case “inside” (Goldie, 2012). Moreover, he admits that this is also the case for *him*, and therefore possibly also for much of the AI research community. This is not only a theoretic tension between two well-understood positions, but affects the very epistemic structure of our academic and engineering disciplines. The “logical” side of this tension is not only ascendant in present society at large, but is more ascendant the more one moves into academic and specifically science/engineering discourse. Having such a tension in the heart of any discipline is problematic, and arguably in the cognitive sciences (and AI) this can be

positively harmful, at least in that it stops us looking in the less supposedly sensible directions. There can be nothing scientific, scholarly, or even sincere about "making believe" on such a grand scale. I have found no AI concepts based these insights by Papert.

Note also that (Winograd, 1991) argues that "[t]*he techniques of artificial intelligence are to the mind what bureaucracy is to human social interaction*". Perhaps it is time to loosen up the stricter aspects of said "*bureaucracy"*. Let us not "*regiment*" the mind, but simulate it more like it is, like it is experienced by us as humans.

-

This thesis explores the intellectual space between Dreyfus (and his successors) and mainstream AI, and between the logical and the subjective. But this is no idle patrol of the intellectual terrain – the worry that AI (as a technology) is stuck is the main motivation. The conclusion will be a concrete recommended direction for AI – a newly-defined method of invention through introspection. In order to concretely demonstrate the novelty and fecundity of the approach, programmed examples are given and discussed (in chapter 7).

1.2 Outlines

This project aims for a third way between the phenomenological critique and classic AI. These scholarly traditions have been antagonistic (see sections 2.1, 2.2, 2.3). Untangling the different stands of existing thought and formalizing this thesis in a linear manner was one of the most challenging aspects of this work. Therefore, several different angles on an overview are useful, in order to serve as a map of the arguments to come.

1.2.1 Scope

If time and word-count were no object, this project would have consisted of three volumes:

1. A proper history of the assumptions underpinning much of AI research.
2. An argument for an alternative, providing new introspection-based avenues for AI development.
3. A thorough set of examples and proper empirical evaluations of the resultant

algorithms.

This thesis as it stands concentrates on the middle part of this enterprise. Not to completely neglect the other parts, a taste of the resultant algorithms with some preliminary empirical information in given in chapter 7, especially sections 7.3 and 7.5. Writing the 1st and 3rd of these volumes is an urgent task, but is out of the scope of a Ph.D. thesis.

1.2.2 Argument outline

The thesis is that "Introspection is recommended for anthropic AI".

Technological AI is distinct from scientific AI. The first aims at producing useful products, the second aims at understanding humans as they are as precisely as possible. This difference calls for different strategies and stresses in these two diverse enterprises (see sections 1.3.2, 1.5). This thesis is about technological AI. Its interest is pragmatic, and short-term: years rather than decades. Since the type of truth required in technology is different than in science (see sections 1.3.2, 1.4), a principal method employed in this thesis is keeping track of the different kinds of truth involved at various junctures.

Human-like AI is distinct from ideal/rational AI (see section 3.2.1). This is like the difference between a human being and some logical-mathematical demon. Most AI has so far aimed at the ideal/rational, perhaps for two reasons: First mathematics is an excellent tool for exploring optimality, and mathematics is readily available – humanity is more complex. Second, especially in looking for technology, the intuitive path is to look for efficient and optimal solutions for concrete problems. However, there are applications where understanding the human way of doing things and fitting in with those customs, sub-optimal as they may be, is precisely what is requisite (see section 3.2.2). Consider a delivery robot – it needs to navigate the inconsistent way that house numbers are occasionally allocated, heed contradictory and partial signs, find the right door without smashing any flowerpots or running over any sleeping cats, understand handwritten notes such as "leave it with the neighbour", etc. A similar conundrum faces a robot dedicated to care for the elderly – we cannot trust an elder to understand how algorithms work – we need the robot to understand (at least partially) the human habits and culture.

Anthropic AI is a new concept (see section 3.2.4). Building a machine that would emulate a full-blown western, modern, well-trained and adult human is a daunting and probably impossible task (at least in the near term). Even if we could produce the ideal (say) Californian, they would be utterly lost in Bangkok. Not only the language is different, but so are many other aspects of the culture. A more promising avenue would be simulating the underlying human intelligence that *allows* one to acquire the diverse cultures. This intelligence, allowing for culture but initially ignorant of any culture, is the aim of anthropic AI. It is contrasted with the "western, modern, well-trained adult". It is trainable to become such a cultured intelligence – but it has the flexibility to become "a Californian" or "a Thai", according to the technological requirement. By emulating the lowest practical level we get a lot of flexibility, and we get much closer to the real human than would a project aiming for "the perfect Californian".

In a sense this research is half-way between proper science and quick-and-dirty technology. The aim is for AI that is not optimal, not getting the task done best-and-fastest (that would be logical/ideal AI), but rather a technology that would do things the human way, which could be slow and accident-prone. On the other hand this does not aim for a perfect or even scientifically correct version of "being human". The motivations of technology demand that we use approximations and deliver a product on an acceptable schedule.

-

In order to explore the different levels of human thought, we cannot afford to shun the **subjective** point of view, nor do we need to (see section 3.3.1).

The most direct access we have to subjectivity is by **introspection**. This raises many objections. The main objection is that introspection is non-scientific, wrong, misleading, or otherwise not a legitimate source of information (sections 4.2.1, 4.2.2, 4.2.3). This originates in (J. B. Watson, 1913). By introducing the distinction between the context of discovery and the context of justification (from philosophy of science) I show that no source of information should be disallowed as a source of ideas (section 4.2.4). Moreover, the level of truth required in technology is less demanding than that required for science (section 4.2.5), so there is no reason to reject *any* source of ideas, *especially* not for technology.

A second objection to introspection-for-AI, somewhat contradicting the first, is that introspection is already in widespread use, and therefore there is nothing new in the above argument. This is counteracted by enumerating the various cases in which introspection was indeed used (section 4.3) and showing that introspection was always used sparingly, shyly, as if there *were* something wrong with it (section 4.6).

Beyond being as legitimate as any other source of ideas, the thesis requires that I show why it should be recommended (not just acceptable), so I need to show that introspection is an effective source of information in some other domain, hopefully a related one. This is done (in section 5.2) by pointing out that introspection is used widely in education. In teaching skills, the teacher could generate their narrative either by recalling verbatim the text used to instruct themselves (decades before) or they could self-observe how they do the skill, and report on that. If the skill is a mental skill, then what we have here is mental self-observation – introspection. Thus the very survival of civilisations for multiple generations stands as testimony that introspective reports somehow carry the gist of how a skill is performed. Introspection is not noise – it carries information of how a skill can be applied – hopefully also by a computer.

The details of how this may be done are covered in chapter 6, including discussion of the kinds of introspection that are more likely to yield interesting new AI. Examples are given in chapter 7. These examples include a novel data type (tracking “trains of thought”), which is further discussed as a basis for future work in chapter 8.

1.2.3 Main antagonists

As is often the case, one's nearest conceptual neighbours are also the antagonists that one critiques most harshly. Here is a list of some of my “nearest neighbours”, a survey of some of their positions is given in chapter 2.

- I shall agree with **Herbert Simon**’s (see sections 2.2 and 4.2) commitment to programming and empirical, pragmatic research. I shall critique him for his lack of imagination, and for his unjustified rejection of subjectivity and introspection (in section 4.2). I concentrate mainly on Simon, but many of his students and successors are very similar in their approach.
- I shall agree with **Hubert Dreyfus** (see sections 2.3 and 4.4) in his commitment

to the subjective point-of-view, and in his general objection to the rationalist views of Simon and the other cognitivists. I shall object to his lack of commitment to programming concrete positive examples, and shall show that programming more phenomenological AI is possible (my examples are found in chapter 7).

- I shall agree with **Winograd & Flores** (see section 2.4) in their recommendation of Gadamer as a possible basis or inspiration for further AI development. I differ with them in their veering away from AI and the lack of any concrete programmed AI examples (a bit like Dreyfus).
- My thinking parallels **Wheeler**'s: I agree with his non-dogmatic pragmatism and with many of his ideas, mainly his notion of action-oriented-representations (see section 8.4.4). We differ in that his scope is broader – Wheeler (2005) deals with "the cognitive world" while I restrict my discussion to the specific field of human-like AI, as a technology.
- I agree with (Brooks, Breazeal, Marjanović, Scassellati, & Williamson, 1999) and others' distinction in human-level intelligence of two levels – the "lower" innate ability which enables the accumulation of culture and skills, and the higher level of a cultured adult. I distinguish anthropic AI (the endeavour of simulating human intelligence *per se)* from any commitment to the specifics of our current, contingent "western, modern, well-trained and adult" ideals about how thinking supposedly *should* be done. I disagree with Brooks et al's contention that the type of intelligence they pre-program can support human behaviour and culture. See also section 3.2.6 about COG and CYC.

This research belongs somewhere between Dreyfus's and Simon's schools of thought, and uses elements of both together with Winograd & Flores's (1986) recommendation of Gadamer as the intellectual context. Having developed some example AI programs, I recognize Wheeler's action-oriented representations in them.

Note that Dreyfus and Simon (and their successors) **talk past each other**: They each "cannot believe" how the other side can be so misguided (McCorduck, 2004, Chapter 9). This has several causes: ontologically Simon (et al) are reductionist physicalists, while Dreyfus is ontologically either an idealist (insofar as he pushes phenomenology in

general) or a Heideggerian (see section 2.5.1). Pragmatically Dreyfus is a philosopher that views his role as writing insightful texts, Simon is an engineer – even a social engineer (see section 2.2). This thesis achieves its goals inside the gap between these thinkers.

1.2.4 Document structure

The first two chapters are introductory – This chapter (introduction) includes some context and preliminary clarifications. A literature review follows in chapter 2.

The next three chapters are the main argument. Chapter 3 gives the entire argument, omitting many details and two main points. The detailed argument on why introspection can have a legitimate role in AI (but has been shunned as if it can not) is in chapter 4. The crucial argument that beyond being acceptable, introspection is to be expected to be positively a good source of ideas for AI, is in chapter 5.

The next two chapters complete the picture: Chapter 6 fills in the details of the recommended methodology for AI development, and discusses some left-over points. Chapter 7 presents working examples of the methodology in action.

Chapter 8 discusses the technical and philosophical consequences of this project, and concludes.

The Appendix (9) is US Patent No. 8,660,670 granted for the main technology presented in chapter 7.

1.2.5 Novelty

This thesis re-evaluates the role of introspection in AI, making use of a variety of perspectives. The main thesis is that "**Introspection is Recommended for Developing Anthropic AI**", i.e. that a development methodology making conscious use of introspection is a promising source of ideas for building the human-like intelligent machines of the future. The main points of novelty are:

1. All human learning, and especially skill learning (section 5.2) is based on a *universal* human foundation. Modelling this foundation in general is the aim of **anthropic AI** (section 3.2). This can be achieved (for AI) by aiming to simulate

in software the human intelligence underpinning but *excluding* any and all contingent culture, especially the western, modern, well-trained and adult type of thinking aimed at by most of AI.

2. Existing approaches to introspection are explored (section 3.3.3), and a conceptual confusion is uncovered, which when cleared up raises an alarming spectre – there can be no such thing as “good” introspection (section 3.3.3.4). The rest of the thesis continues with the understanding that we will always deal with “bad” introspection.

3. The role of introspection in cognitive science and specifically in AI research is explored in detail (chapter 4), and shown to have never been wholeheartedly adopted in AI (sections 4.3, 4.6). Exploring the reasons given for shunning introspection, these do not stand to scrutiny, especially in the context of AI as a technology (section 4.2).

4. Introspection is shown to be an important avenue for the transmittal of mental skills from one generation to the next in humans, and hence cannot be “nonsense” or “noise” but is rather an efficacious source of information on how human skills work (section 5.2). Since the transferral of human skills to a computer (rather than to another human) is the very essence of the process of developing human-like AI, introspection is recommended as a source of ideas for designs.

   In other words: if a teacher uses introspection to extract mental skills from themselves to transmit them to a student, an AI developer can also use introspection to uncover the human skills that they want to transfer to a computer.

5. AI practitioners generally shunned subjectivity (section 4.2), but produced working systems, while critics of AI (mainly Dreyfus) extolled subjectivity, but did not produce any such systems (section 4.4.3). In their writings, these two communities talk past each other (chapter 2). In providing a working example of AI based on introspection I provide a concrete manifestation of a compromise between these camps: programmed subjectivity, through introspection (chapter 7).

6. This program implements a data type that (to the best of my knowledge) is unlike any that has been used in AI before (section 7.5). Further evidence of its novelty is provided by the appendix, a US Patent on this technology.

7. From a continental perspective, this data type provides the first concrete AI design that can be called “Gadamerian”, as recommended by (Winograd & Flores, 1986). This is a novel step in analysing what “Heideggerian AI” could be (Dreyfus, 2007), by isolating the hermeneutic aspect of Heidegger's philosophy (as detailed by Gadamer) from the rest of Heidegger's philosophy (section 8.4.2).

Some may view this thesis, “Introspection is recommended for developing anthropic AI” as being weak or blurry. This critique, far from pointing out a fault in the project, points to some main contributions: The mainstream AI community has been ignoring the subjective while the mainstream critics (Dreyfus) have spoken of little other than subjectivity (point 5 above). Both made the same mistake of treating AI as a single project, to be examined from a single perspective. In delineating anthropic AI as separate from at least two other categories (point 1 above), and in showing that subjectivity (in the form of introspection) is useful for producing concrete AI systems (point 4 above), this thesis not only provides a more advanced analysis of the field(s) of AI, but also makes a first programmed contribution using full-blooded subjective methods (points 6-7).

1.3 The specific field of this thesis

Before the literature review, we need to clarify some points regarding the specific field of this thesis, and how different disciplines use different notions of truth. This will also be a principal distinction in the rest of the thesis.

1.3.1 Philosophy of AI

This thesis is in the field of philosophy of AI. Being a relatively small field, this calls for a discussion of the very nature of the field.

One way of looking at philosophy in general is as the all-encompassing love of wisdom.

The **specialised fields** (physics, medicine, architecture) can be seen as spheres of knowledge from which philosophy-as-such has to a large degree retreated. Questions of

the interaction of materials (for example) are no longer examined by philosophy, or even “natural philosophy”, but are examined by chemistry and physics. So extending the graphic metaphor of “spheres” for a moment, philosophy (as a field) is left responsible for a “Swiss cheese” - all that is left of knowledge after the specialists took charge of their spheres – the areas between fields, and the areas far removed from any specialised field of research.

But this “subcontracting” of the specific spheres of knowledge to their respective experts is not total. Difficult questions such as those raised by quantum mechanics are discussed often by philosophers (Ismael, 2015). Philosophers also reserve the right to discuss the “meta” fields, such as philosophy-of-science, ethics-in-medicine, etc. In fact, any questions that are seen as neglected within any of these specialised fields can be taken up by a philosopher. In the case of AI, Dreyfus stands out as a philosopher who intervened forcefully in a field in which he was not part of the research community. Moreover, issues of methodology are often seen as “philosophical” even by practitioners of the specialised fields who would otherwise be disinterested or even hostile to philosophy as such.

The field of this thesis is philosophy-of-AI *as a technology* in a sense to be clarified through the rest of this thesis. The concern is with the *methodology* of AI invention and development – with how AI researchers come up with their ideas.

AI may be viewed as “*the intellectual core of orthodox cognitive science*” (Wheeler, 2005, p. 68), or it may be viewed as a technological field (S. Russell & Norvig, 2013). This thesis is interested in AI-as-technology. Any technological models may optionally later inspire or form the basis for theories in psychology, but that is not the focus here. More detail of this is found in sections 1.5 and 8.3.1.

It seems that “philosophy of AI” (at least AI as a technology, which is the concern here) would be a sub-field of “philosophy of technology”. Again, this warrants a closer look.

1.3.2 Philosophy of technology

(Franssen, Lokhorst, & van de Poel, 2013, sec. 2.2) divide theories of technology in terms of their position on the question of whether technology is merely “applied science” or has some different inherent nature.

On the side that technology *is* applied science Franssen et al quote Bunge saying that "*technology is about action, but an action heavily underpinned by theory—that is what distinguishes technology from the arts and crafts and puts it on a par with science*". Bunge seems to be stressing the distinction between the (old) crafts like carpentry and modern technology, with its heavy reliance on science (e.g. computers). Computers and many other enabling technologies of our current lifestyle would be impossible without quantum mechanics. This break between the old crafts and the (threatening) new technology is important to ethicists who worry about new technologies, not least the later Heidegger (2009).

On the other side we find both Skolimowski and Herbert Simon (see section 2.2) who see continuity between the old crafts and modern technology. Skolimowski says that "*science concerns itself with what is, whereas technology concerns itself with what is to be*". On the other hand (an earlier version of) (Simon, 1996b) says that "*the scientist is concerned with how things are but the engineer with how things ought to be*" (Franssen et al., 2013). In terms of their characterisation of science there is little difference between Skolimowski's *"what is"* and Simon's *"how things are"*. In terms of technology, Skolimowski's *"what is to be"* is quite different from Simon's *"how things ought to be"*, at least in terms of approach:

Skolimowski's position, if taken to mean "what it is to be" in terms of physics or metaphysics seems far less interesting *in terms of technology* then if we read it as a call for a functionalist definition of artefacts: "What *would it take* for something to become X". A car *is* a horseless carriage in terms of the functions it aims to fulfil, but not in terms of the internal materials and techniques involved. So by this (functionalist) reading of Skolimowski he is saying that technology is about the internal functions that give rise to the overall functions, and presumably technologists are engaged in fulfilling these functions in whatever way feasible. Simon's definition ("*how things ought to be*") seems to be either teleological, or stressing the external functions of any contraption. Under the later reading Simon is more interested in how a contraption functions *for us and our larger purposes* than in its internal mechanics. This is in line with Simon's interest in the social sciences, public and business administration, public policy, etc. (see section 2.2). So the difference between Simon and Skolimowski seems to be one of stress on the outer vs. inner functioning of artefacts (respectively).

In these definitions we see that the crux of the matter for technology is function rather than truth. Truth is for scientists, function is for technologists. This is a key point in this thesis.

As a consequence of this difference in relation to truth there are also the following differences:

- Science has an insatiable appetite for precision. Technology always has a practical limit to any obsession with precision – a "tolerance".
- Science aspires to generalize, and come up with a definitive theory of everything – one ultimate reality, one unified theory (This is reminiscent of monotheism, see below). Technology on the other hand deals with multiple perspectives: The perspectives of the various parts of a machine, and the thermal, mechanical, and electric perspectives, etc. of each part and/or the whole.

A key example of the use I will make of the lower truth-requirement in technology is in sections 4.2.2, 6.3.5.1: If, as (Nisbett & Wilson, 1977) show, introspection gives us only the outline of *what* is accomplished by the mind without the *how*, in AI programming (as a technology) we can substitute whatever technical trick we have (in our skills as programmers) to achieve something similar in a computer. In psychology, as a science, this of course would not do – scientists demand (ultimately) an explanation of mental processes in terms of the brain, and physics.

1.4 Notions of truth

It is the aspiration of science to aim for the most accurate facts, and to stick to the truth with great zeal. This commitment to strictly follow the best version of the truth leads, amongst other things, to disdain for the humanities, or any of the "loose" sciences. In this sense, it is the main project of Psychology (Costall, 2006; J. B. Watson, 1913) to move from the "loose" or "woolly" sciences nearer the more "proper" sciences. Mathematics, sometimes called "the queen of sciences", has little content *but* the demand to stick 100% to a well-defined notion of truth. This thesis will show that in technology not only are we not obliged to such a strict adherence to the truth, but we are positively hindered by such a zealous loyalty to this notion of maximal truth. In a sense the main methodology of this thesis is examining different types of truth-claims, and

keeping these truth-types distinct and clear.

1.4.1 The idea of a single truth

Many people believe that there is one truth. This idea originates in monotheism, but has mutated into the scientific world, not least in early 19th century France.

Auguste Comte (b 1798 d 1857) was a French intellectual trying to make sense of the world from a post-revolutionary perspective after the traditional institutions of monarchy and church have lost their meaning. He created a doctrine called positivism (Bourdeau, 2014; Mill, 2013), that held that all sciences were going to eventually be unified in a single logical structure with mathematics at the base and physics, chemistry, biology etc. leading up to a full understanding of individual humans, and eventually societies. His movement had many adherents, not least John Stewart Mill, and the designers of the Brazilian flag who put "order and progress" - Comte's slogan – on the flag. Positivism, believing in inexorable human progress, did not survive the horrors of WW I. However, it partially reincarnated in logical positivism (the Vienna circle), and the hope that the sciences will coalesce and ultimately leave no empty space for ignorance between them held on to a large degree. Note, however, that (J. B. Watson, 1913) was before the war (which started July 1914), and that the founding of behaviourism was motivated by his need to bring humans and other animals under the same scientific field.

Regardless of such speculations about the origin of the idea, we can see in today's literature a commitment to a single truth, be it a scientific-administrative physicalist truth (Simon, see section 2.2) or a phenomenological, idealist or Heideggerian truth (Dreyfus, see section 2.3). This yearning for a single truth is expressed also by attempts to solve the mind/brain problem once and for all, such as the Blue Brain project (Markram, 2006). I propose, for AI, a more short-term and practical alternative, using multiple and competing perspectives concurrently. This is a bit like Minsky's (1991) "scruffy" notion.

1.4.2 Perspectivism

For AI as a technology, the discussion in section 1.3.2 of technology vs. science brings us to consider perspectivism, which allows us to hold both horns in case of a dilemma, understand several contradictory aspects, and determine what kind of truth is requisite at

any moment. In a sense a central aim of this thesis is to show how wrong types of truth (wrong perspectives) were sometimes used in AI research, and a way of rectifying that.

For now, let us look at perspectivism per se:

> *Against positivism, which halts at phenomena--"There are only facts"--I would say: No, facts is precisely what there is not, only interpretations. We cannot establish any fact "in itself": perhaps it is folly to want to do such a thing [...]*
>
> *In so far as the word "knowledge" has any meaning, the world is knowable; but it is interpretable otherwise, it has no meaning behind it, but countless meanings. –"Perspectivism."*
>
> *It is our needs that interpret the world; our drives and their For and Against. Every drive is a kind of lust to rule; each one has its perspective that it would like to compel all the other drives to accept as a norm. (Nietzsche, 1889, sec. 481)*

As an example, consider an armchair in the social room in our university department. What is it? Which of the following is “The Truth”, or “Reality”? It could be:

- a chair, an armchair
- part of the equipment of the social room/the department/the university/the educational system/the UK/the west/humanity/an elitist plot to exclude the uneducated/etc.

But it can also be seen as:

- a physical thing, a solid, with a location/certain weight and size/existing in time
- pieces of wood and cloth, arranged a certain way
- a large collection of dead (mainly plant) cells
- molecules; elements; atoms; sub-atomic particles
- quarks or whatever else the physicists will come up with in the future[1].

---

1 An aside on the ontology of recent generations: It seems that (for some) the whole discussion of “what is”, or “what is real” has, for the first time in the history of philosophy, been subcontracted out to some other discipline, namely physics. Even when we treat occasionally “atoms”, “sub atomic particles”, “quarks” or suchlike as the building blocks of the objective universe, we mostly agree that if physics found some new subdivision (below the current “standard model”) from which all quarks, leptons, bosons etc. are made, we would immediately accept any new scientific consensus as our new ontology.

But it is also:

- a coloured object; mostly greyish-blue; part of colour scheme; part of a setting
- old; discoloured; damaged; dangerous; a health-and safety violation
- a disgrace to the department (and all other bodies up to Humanity, see above)
- a relic from a bygone era.

And we could go on endlessly, describing every part in every context and from every perspective.

This multitude of perspectives is visible in everything about us, and we choose which perspective to use, as appropriate, while we interpret the world. If we were to look at a person rather than a chair, the possibilities of interpretation would be even greater. None of these perspectives is more true or real than any other – however each may be considered more appropriate for a given context or perspective.

Note that often people tell each other to "get real" or to discuss "**the real world**". But in what sense can we say that some perspective is more real than another? To return to our example, for a lawyer, the chair is a health-and-safety violation, a liability, a risk. For the biologist it is dead organic matter. For me (on some days) it is ugly, or (on other days) homely. How can we decide this? I propose that we *don't*, and that we suspend any discussion of one objective reality, and treat any demand to "get real" as suspect, even violent attempt to impose a specific perspective (see the later part of Nietzsche's quote above).

1.4.3 Perspectives, realities, agendas, Occam

This particular section deals with motivations for different ways of thinking. Accepting or rejecting any of the statements here does not change the main argument. It is provided as a guide as to why many people reject perspectivism (in practice if not explicitly). Presenting the extent of our "one-truth" prejudice with proper rigour is outside the scope of this thesis.

There are two main motivations to collapsing a multi-perspective discussion into a single perspective:

- The demand to narrow down a discussion to a single perspective is driven by an active agenda – if there were no agenda there would be no need to focus. This is true both of the “calendar” meaning of the word “agenda” – “we need to get something done in limited time”, and of the more sinister-sounding “political” agenda, where there is a group specifically interested in shutting down some discussion by making it “unreal”. An example “close to home” is how Watson shut down introspection (Costall, 2006), see section 4.2.1 and (the end of) Nietzsche’s quote above (section 1.4.2).
- The other motivation for collapsing a multi-perspective discussion into a single perspective is individual. Having a single perspective simplifies a discussion greatly, and allows faster progress (at the cost of depth). Moreover, (looking also from a child's mentality) completely believing in a single truth gives one a sense of security and closure, as a child has once an adult tells them “everything is OK”. Having such closure is consoling, and allows us to function in a world that is inherently unpredictable and frightening. This also explains the clamour for strong leadership in times of insecurity, regardless of the quality of such leadership (Fromm, 2011). On a collective level, this fits with our human tendency to follow miracle-workers. In the insecurity of not knowing which perspective to adopt, we clamour for a wise or strong person to lead us, to absolve us from the need to worry about the various perspectives ourselves. Examples of such miracle-workers range from Moses and Jesus to the USA scientific-military-industrial complex and its most impressive spectacles – the detonation of atomic bombs and space travel. The more sinister cases of humanity's hankering after simplicity are those of demagogue-dictators, as discussed by (Fromm, 2011).

As we have seen, we often want one powerful truth, so we can become its loyal slaves and always win, by this one truth winning. Occam's razor is the most prominent tool for reducing multiple truths into one, and not surprisingly is a rare relic from the centuries before the scientific revolution that is still revered as if the pope still instructs us as to what to believe.

So let us, for now, eschew any hard-and-fast discussion of one "reality" and just examine perspectives. This of course brings up the question: in what perspective, in which sense, do I make the claims of this thesis?

1.4.4 In what sense is this thesis true?

In presenting this thesis, I claim, at least in some sense, that it is true. But what kind of truth do I claim for this thesis? I claim a pragmatic truth – as befits a thesis that is ultimately about technology. So why perspectivism? Because adopting different perspectives at different times *works*. Engineers do it all the time when they design modules – they design the interaction between the complete modules, and also the internal structure of each module. They design the electronic characteristics of a system, and then move to look at the thermal design, and then the mechanical structure, moving between these distinct perspectives. In a sense this thesis will promote perspectivism as an (analytical) tool, and pragmatism as value-system (for technology development). Later (especially sections 3.3, 4.7, chapter 6) I will discuss subjectivity and introspection, introducing many perspectives and claiming (some) validity for some of them. This thesis is about a method for developing ideas; ideas for AI. But ideas as such are not the bottom line – the bottom line (in technology) is letting what works win, without prejudice towards any particular perspectives.

So what needs to be shown is that this thesis points to plausibly profitable avenues in anthropic-AI research. I need to show how these profitable avenues of research are neglected by current conceptions, and why these avenues make promising starts at addressing interesting fields.

1.4.5 Notions of truth: summary

A central theme, perhaps the central method running throughout this thesis is how different notions of truth are used in different contexts in AI research, and how some of these may need to be re-examined. The demand for less truth than necessary will lead us to absurdities, but the demand for more precision than can be found constricts our ability to invent. As Aristotle had it: "*It is the mark of an educated man to look for precision in each class of things just so far as the nature of the subject admits*" (Aristotle, 2009).

1.5 Science vs technology and human-like vs rational

The following two distinctions are important to be born in mind while discussing AI, in terms of the type of motivation involved:

- Some researchers build AI models in order to *scientifically understand* humans, mice, insects, neural functioning or some other scientific questions (usually long term), while others build AI in order to solve problems of *technology* in the short term (the contrast between science and technology as motivation should not be confused with the contrast between them as to the types of truth required, see section 1.3.2).

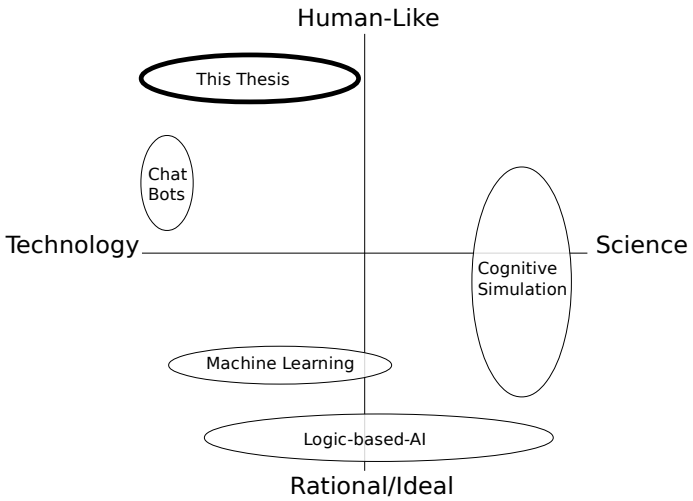


Illustration 1.1: Locating different AI fields in key distinctions

- Some researchers aim for Human-like AI, and some aim for rational or ideal AI (see section 3.2.1).

These two distinctions should be seen not as dichotomous, but as continuous.

For examples, referring to Illustration 1.1, see the technological motivation of Chat-Bots, from Eliza (Weizenbaum, 1966), to commercial chat-bots and the entrants of the Loebner prize (Mladenić & Bradeško, 2012). These aim for human-like performance here-and-now (technology), using any trick available. Meanwhile for Cognitive Simulations (or computational psychology) (Sun, 2008) the aim is to simulate for science working models of human cognitive faculties. In sharp contrast, search engines (such as Google) aim at the best possible result regardless of how humans would fare at the same task, as do other Machine Learning implementations. Logic based AI is an attempt to explain humans scientifically using various types of logic (Bringsjord, 2008), and has also been the basis of much of ("good old fashioned") classic AI, including its

large technological component (McCorduck, 2004).

Both these distinctions are a bit more complex than they seem, at least for AI: We *do*, at the end of the day, need to translate any human-like ideas into a formal computer language (see section 6.3.5) in order to make an AI system work – such is the nature of the technology available to us at the moment. More detail on what is meant by "human-like" and what I aim for are in section 3.2.

The distinction between technology and science isn't sharp either – a working system demonstrates to science what can be achieved by the underlying mechanism, and conversely scientific models (such as neural models) serve as the bases (or inspirations, see sections 3.1, 6.1.1) for AI systems. A demonstration of how seriously this link is taken by contemporary science could be seen in a note on Richard Feynman's blackboard at the time of his death, reading "*what I cannot create, I do not understand*"[2] (Feynman, 1988; Resnick, 1993). Another point to recall in the interaction between science and technology is that in every generation people conceive of humans in terms of the latest technology, e.g. Aristotle's "clay vessel with a divine spark", notions of the heart as a furnace, Freud's hydraulic models and today's notions of the brain as computer (Bolter, 1984).

So the project presented here is about human-like, technological AI. An important part of the methodology is keeping track of truth-claims, and what kinds of truth are being claimed.

2 Photo of blackboard: http://archives.caltech.edu/pictures/1.10-29.jpg

2 Literature review

Table of Contents

2 Literature review....23
2.1 The Cognition-vs-Phenomenology debate....24
2.2 Simon....28
2.2.1 Economics....29
2.2.2 Hostility to subjectivity - rationalistic....29
2.2.3 AI....30
2.3 Dreyfus....31
2.3.1 Part I....31
2.3.2 Part II – Assumptions Underlying Persistent Optimism....32
2.3.3 Part III – Alternatives to the Traditional Assumptions....34
2.3.4 Dreyfus's current position....37
2.4 Winograd & Flores....38
2.4.1 Understanding and being....39
2.4.2 Cognition as a biological phenomenon....42
2.4.3 Language as listening and commitment....42
2.5 Hermeneutics and Gadamer....43
2.5.1 Phenomenology, hermeneutics, and other disciplines....43
2.5.2 The hermeneutics of Heidegger and Gadamer....45
2.6 Literature review: summary....47

This thesis aims to establish that “introspection is recommended for anthropic AI”. In so doing, a development paradigm for AI will be explored (approximately) half way between classic AI and the phenomenological critique of AI as initiated by (Dreyfus, 1979). The literature review will start with understanding the two camps – cogitivism and classic AI on one side, and phenomenology and its critique of AI on the other.

Section 2.1 focuses on the heart of the disagreements between cognitivism and phenomenology.

Section 2.2 Presents classic AI through the central figure of Herbert Simon. Simon is the pre-eminent thinker in classic AI, the only person ever to receive both the Nobel prize and the Turing prize.

Section 2.3 Presents Dreyfus and his critique of AI. Dreyfus (1979) is “the voice of one crying in the wilderness: 'Make clear the way for Heidegger!'…”, arguing against a

tradition in AI that Winograd & Flores will later (1986) call the rationalistic tradition. Sadly, Dreyfus has no concrete, coded contribution.

Section 2.4 Presents Winograd & Flores, who aim to bridge this divide and make some more concrete progress. They come up with ideas for a working system, but that system starts a new field of technology called “group-ware” rather than contributing to AI as such. One of the contributions of their analysis is adding the hermeneutics of Gadamer to the discussion.

Section 2.5 surveys hermeneutics and Gadamer in more depth.

In a nutshell, my position (that introspection is recommended for anthropic AI) will be a compromise between Simon and Dreyfus, and can be seen as a continuation of Winograd & Flores's work, especially their suggestion of Gadamer. Unlike Dreyfus, this thesis aims to be positive, to show what *can* be done based on a more critical approach to AI. Unlike Simon it takes account of the subjective point of view. Like Winograd and Flores, it will look at a broader critique of AI than Dreyfus did, and will end up looking surprisingly like the work of Gadamer (1976, 1979). Unlike Winograd & Flores it will stay in the field of AI.

In introducing the background of this thesis, more space will be dedicated in this chapter to continental concepts, as they would be more foreign to some readers than the analytical terminology.

2.1 The Cognition-vs-Phenomenology debate

This thesis proposes a novel avenue in AI development (introspection), which is between the classic-AI cognitive tradition and Dreyfus's critique of AI. Therefore, as the first order-of-business it would be useful to contrast the cognitivist position with Dreyfus's starkly and clearly, using two founding texts. We will find that these two camps talk past each other: cognitivism is essentially reductionist-materialist, while phenomenology has either an idealist or Heideggerian ontology. Cognitivism understands the mind “from the outside”, as an engineer would, while phenomenology is interested in “what it is like” to be/have a mind. Sociologically, phenomenology is heavily based on a German-language literary tradition, while cognitivism is heavily based on the metaphor of mind as machine.

The book “Cognitive Psychology” (Neisser, 1967) coined the term “cognitive psychology” which gave the name “cognitive” to the later “cognitive science” (Boden, 2008, p. 16). In the first paragraphs of Neisser's book - arguably the defining part of the defining text of the entire field – he says:

> *It has been said that beauty is in the eye of the beholder. As a hypothesis about localization of function, the statement is not quite right—the brain and not the eye is surely the most important organ involved. Nevertheless it points clearly enough toward the central problem of cognition. Whether beautiful or ugly or just conveniently at hand, the world of experience is produced by the man who experiences it.*
>
> *This is not the attitude of a skeptic, only of a psychologist. There certainly is a real world of trees and people and cars and even books, and it has a great deal to do with our experiences of these objects. However, we have no direct, immediate access to the world, nor to any of its properties. The ancient theory of eidola, which supposed that faint copies of objects can enter the mind directly, must be rejected. Whatever we know about reality has been mediated, not only by the organs of sense but by complex systems which interpret and reinterpret sensory information. The activity of the cognitive systems results in—and is integrated with—the activity of muscles and glands that we call "behavior." It is also partially—very partially—reflected in those private experiences of seeing, hearing, imagining, and thinking to which verbal descriptions never do full justice.*
>
> *Physically, this page is an array of small mounds of ink, lying in certain positions on the more highly reflective surface of the paper. It is this physical page which Koffka (1935) and others would have called the "distal stimulus," and from which the reader is hopefully acquiring some information. But the sensory input is not the page itself; it is a pattern of light rays, originating in the sun or in some artificial source, that are reflected from the page and happen to reach the eye. Suitably focused by the lens and other ocular apparatus, the rays fall on the sensitive retina, where they can initiate the neural processes that eventually lead to seeing and reading and remembering. These patterns of light at the retina are the so-called "proximal stimuli." They are not the least bit like eidola. One-sided in their perspective, shifting radically several times each second, unique and novel at every moment, the proximal stimuli bear little resemblance to either the real object that gave rise to them or to the object of experience that the perceiver will construct as a result...'*

So we see that Neisser accepts the following perspectives (see section 1.4.2 about perspectives) as legitimate:

- A person is an animal, with their brain, sense organs etc.
- with a “complex” “cognitive system”, inside
- in a “real world”, with
    - “trees and people and cars and even books”, and
    - “small mounds of ink” that generate
- “patterns of light at the retina” - a process rather than a thing.

And Neisser rejects “eidola”. “Private experiences” are only accepted as end-results, not as having any causal efficacy, beyond being described. It seems that Neisser wants to reduce the number of elements in his world-view as much as possible - he is trying to make his world-view simpler. This could have two motivations: either to make a technology based on this model more feasible or because he has inherited an aversion to complex explanations from his predecessors, be they Occam, Watson (see section 3.3.3), or the other early cognitivists (Pear, 2007, pp. 111–115).

Dreyfus, on the other hand (1979, p. 269), will have none of this. Quoting from Neisser's passage above:

> *There is certainly a real world of trees and people and cars and even books… However, we have no direct, immediate access to the world, nor to any of its properties.*

Dreyfus says:

> *Here* [...] *the damage is already done. There is indeed a world to which we have no immediate access. We do not directly perceive the world of atoms and electromagnetic waves (if it even makes sense to speak of perceiving them) but the world of cars and books is just the world we do directly experience. ... we saw that at this point, Neisser has recourse to an unjustified theory that we perceive "snapshots" or sense data. His further account only compounds the confusion:*
>
> > *Physically, this page is an array of small mounds of ink, lying in certain positions on the more highly reflective surface of the paper.*
>
> *But physically, what is there are atoms in motion, not paper and small*

> *mounds of ink. Paper and small mounds of ink are elements in the human world. Neisser, however, is trying to look at them in a special way, as if he were a savage, a Martian,* ***or a computer****, who didn't know what they were for. There is no reason to suppose that these strangely isolated objects are what men directly perceive (although one may perhaps approximate this experience in the very special detached attitude which comes over a cognitive psychologist sitting down to write a book). What we normally perceive is a printed page.*
>
> *Again Neisser's middle-world, which is neither the world of physics nor the human* [experience] *world, turns out to be an artefact. No man has ever seen such an eerie world; and no physicist has any place for it in his system. Once we postulate it, however, it follows inevitably that the human world will somehow have to be reconstructed out of these fragments.*
>
> > *One-sided in their perspective, shifting radically several times each second, unique and novel at every moment, the proximal stimuli bear little resemblance to either the real object that gave rise to them or to the object of experience that the perceiver will construct as a result.*
>
> *But this whole construction process is superfluous. It is described in terms which make sense only if we think of man as a* ***computer*** *receiving isolated facts from a world in which it has no purposes; programmed to use them, plus a lot of other meaningless data it has accumulated or been given, to make some sort of sense (whatever that might mean) out of what is going on around it.*
>
> *There is no reason to suppose that a normal human being has this problem, although some aphasics do. A normal person experiences the objects of the world as already interrelated and full of meaning. There is no justification for the assumption that we first experience isolated facts, or snapshots of facts, or momentary views of snapshots of isolated facts, and then give them significance. The analytical superfluousness of such a process is what contemporary philosophers such as Heidegger and Wittgenstein are trying to point out. To put this in terms of Neisser's discussion as nearly as sense will allow, we would have to say: "The human world is the mind's model of the physical world." But then there is no point in saying it is "in the mind," and no point in inventing a third world between the physical and the human world which is an arbitrarily impoverished version of the world in which we live, out of which this world has to be built up again. (*stress added)

So Dreyfus rejects the “ink mounds” and extols phenomenology (human experience of a world full of meaningful things and situations). Phenomenology (see section 3.3.1.5) is systematic peer-reviewed introspection (Gallagher & Zahavi, 2012, pp. 28–29). One could be fully sympathetic with Dreyfus, but it is important to also understand Neisser -

he goes to the “ink mounds” level probably because he has an eye out for the scientific-reductionist or *engineering* perspective. The ink mounds are not only some “Martian” invention of cognitive scientists, but also an approximation of what a camera would see as pixels. The ink mounds are also what an engineer would need in order to *replicate* a page. And replication, in the 20th century, is considered a hallmark of understanding: as mentioned in section 1.5, on the death of Richard Feynman this text was found on his office blackboard: “*What I cannot create, I do not understand*” (Feynman, 1988).

Arguably Dreyfus is the more thoroughly philosophical (or more dogmatic) of the two, trying to distinguish “real” categories, while Neisser is more pragmatic and technological.

Let's now move to a more systematic presentation of Simon & Dreyfus's positions. More on how they talk past each other is found in section 1.2.3.

2.2 Simon

Herbert Simon (b. 1916, d. 2001), was one of the most prolific thinkers of the 20th century. His contributions span Public Administration, Business Studies, Psychology (where he is considered one of the instigators of the cognitive revolution (Pear, 2007, p. 113)), Economics, Operations research, Mathematics, Statistics, Computing, and Artificial Intelligence. Indeed, his “*more than 900 publications*” span in addition to computers and AI “*every social science discipline other than anthropology*” (Augier & March, 2001). “*As much as any one person, Herbert A. Simon has shaped the intellectual agenda of the human and social sciences in the second half of the twentieth century*” (Turkle, 1991). Specifically in AI, arguably over half of the achievements of GOFAI were his own or his students'. He is the only person (so far) to win both the Turing prize (1975) and the Nobel prize (in economics, 1978). He continued Watson's objections to introspection (J. B. Watson, 1913, 1920), and took part in bringing much of behaviourism’s heritage into the cognitive fold (Costall, 2006). For all his breadth, he saw himself as “*a monomaniac. All my life I have been studying one thing: human decision making*” (Feigenbaum, 1989). His impact was not accidental. His energy and conviction showed him in the light of “*a proper missionary*” (Augier & March, 2001).

2.2.1 Economics

Simon viewed as his main contribution (in the field economics) the notion of **bounded rationality** (Simon, 1996a, p. 165). The idea of bounded rationality was a rebellion against classical economics, and that theory's faith in "*'economic man', who in the course of being 'economic' is also 'rational' "* (Simon, 1955). Simon's contention was that the circumstances in which a human agent finds himself place restrictions on the human's rationality, and therefore the idea of an optimising, totally rational "economic man" is unrealistically optimistic; humans are not as rational or well informed as classical economics would have it. Human rationality is bounded by the information available and the amount of time and resources one can practically devote to any decision. Similarly in AI, Simon propounds "satisficing", which is his term for a solution that is not optimal, but "good enough", or "fit for purpose". He later related satisficing to the notion of heuristics in AI.

2.2.2 Hostility to subjectivity - rationalistic

Simon was a scientist of the rationalistic tradition through and through (as defined by (Winograd & Flores, 1986), see section 2.4). His basic world-view was that humans are rational, his main metaphors for life were a maze, or a chess game (Simon, 1996a, p. 113). "*For me mathematics has always been a language of thought. I don't know precisely what I mean by that... Mathematics – this sort of non-verbal thinking – is my language of discovery*" (*Ibid.* p. 106).

He was uncomfortable with subjectivity and often distanced himself from it - in himself and others. In his autobiography (Simon, 1996a), he only refers to himself in the first person once he is grown up – the vicissitudes of a child's emotions were a bit too much for him to identify with or discuss in later life. In his autobiography, when describing his early years he refers to himself in the third person, as "*the boy*". Moreover, in his uncomfortable relationship with all things subjective, he was not versed even in the everyday terminologies of subjectivity, confounding "introspection" with "introversion": "*the boy himself was incorrigibly introspective*" (*Ibid.,* p. 19).

Simon bought into Watson's ambition (Costall, 2006) that psychology should be a science on a par with chemistry and physics. However he falls into hubris, and takes his ambition to be already substantially fulfilled, making his own projects into the building

blocks of his world-view, with faultless circular logic (Costall, 2006; Dreyfus, 1979): "*If chess plays the role in cognitive research that Drosophila does in genetics, the Towers of Hanoi is the analogue of E Coli...*" (Simon, 1996a, p. 327).

2.2.3 AI

The issue of subjectivity and introspection does not explicitly show up in Simon's writings about AI. Rationalistic symbol-manipulating AI was (for him) the future: "... *enlarging symbol manipulation to embrace much more than deductive logic. Symbols can be used for everyday thinking, for metaphorical thinking, even for 'illogical' thinking. This crucial generalisation began to emerge at about the time of world war II, though it took the appearance of the modern computer to perfect it*" (*Ibid.* p. 193). Throughout his AI career he was "... *interested in simulating human problem solving, and not simply demonstrating how computers can solve hard problems*" (Simon, 1996a, p. 209). So much so that he explicitly claims in a 1956 letter to Russell that his software proves theorems like a human does (*Ibid.* p. 207, see also pp. 234, 274, 331). And yet, there are no signs of subjectivity or fallibility, beyond his own theory of "bounded rationality".

Simon "*saw problems as generally decomposable into hierarchical structures*" (Augier & March, 2001; Simon, 1989). This view is visible in his AI efforts, such as the General Problem Solver, and the Logic Theorist. His faith in decomposability, and in the validity of general solutions for wide areas shows also in his contention (above) of being only interested in one question, human decision making, while publishing research in many different fields.

Simon towered over the field of AI also in terms of methodology, and in terms of how the field viewed itself: Simon viewed AI as a science (Simon, 1996b), and as inseparable from his other fields of study – in this sense he pre-dated Wheeler's (2005) definition of AI as the "intellectual heart of cognitive science".

In terms of methodology, Simon (with his usual AI collaborator, Newell) established his ideas of the rules of the field. As (Feigenbaum, 1989, p. 10) has it:

> *The work of Newell and Simon impressed upon AI the methodological paradigm of empirical science. The methodology dictates this 'scientific loop':*

a. design … based on the information processing model

b. test … based on computer programs you write to represent your design

c. measure … based on actual computer runs of these programs (not 'pencil-and-paper', not 'armchair thinking', not 'theorems')

d. redesign… based upon the discoveries made about the behaviour of the modelled

This is the 'coin of the realm', as Newell labelled it.…..

More details of his work are mentioned as needed in the succeeding chapters, especially sections 3.3.3, 4.2.2, 4.3.3.

2.3 Dreyfus

Dreyfus's book, “What computers can't do” (1979) may have been better titled “What cannot be formalised”. Beyond being quite polemic against the AI community, he exposes their over-optimism that he blames on wild extrapolations, and shows their underlying assumptions that he finds questionable. He presents very sketchily an alternative based on the works of Heidegger and Merleau-Ponty, but this alternative does not lead to anything that can be formalised and programmed.

The rest of this section is a summary of “What Computers Can't Do” (Dreyfus, 1979) and his more recent published position (Dreyfus, 2007).

2.3.1 Part I

To Dreyfus, the assumptions of AI start with Plato (Dreyfus, 1979, pp. 67–68), who looks in Euthyphro for the “necessary and sufficient” conditions for piety. To him, this is the first quest for “effective computation” - a blind procedure that would allow a conclusion to be reached with no human discretion. He skims over Aristotle, and quotes Hobbes' assertion that “*reason is nothing but reckoning*” (*Ibid*. p. 69). Leibniz saw his algebra as a means to calculating the “characteristics” of things, perhaps the first allusion to symbols as a basis for AI. Dreyfus moves along swiftly to Boole, Babbage, and Turing (*Ibid.* p 71).

Dreyfus then gives a history of AI, and turns to the wild optimism of early years, including Simon's 1956 predictions (Dreyfus, 1979, pp. 81–82):

> *1. That within ten years a digital computer will be the world chess champion, unless the rules bar it from competition.*
>
> *2. That within ten years a digital computer will discover and prove an important new mathematical theorem.*
>
> *3. That within ten years most theories in psychology will take the form of computer programs, or of qualitative statements about the characteristics of computer programs.*

Dreyfus then launches into a detailed debunking of these predictions.

Next (*Ibid.* pp. 91-100) Dreyfus surveys AI work from 1957 to 1962 in what he terms "**cognitive simulation**", including language translation, problem solving and pattern recognition. He shows that the efforts in these fields were characterised by early success, enthusiasm, failure, and sometimes pessimism and sometimes outright denial of the difficulties.

His diagnosis of cognitive simulation is that this project has been on the wrong side of four distinctions:

- Fringe consciousness vs Heuristically guided search (*Ibid.* p. 100)
- Ambiguity tolerance vs context-free precision (*Ibid.* p. 107)
- Essential/inessential discrimination vs trial-and-error search (*Ibid.* p. 112)
- Perspicuous grouping vs character lists (*Ibid.* p. 120)

The next phase of AI was what Dreyfus calls "**semantic information processing**": Bobrow's "student" (*Ibid*. p. 132), Evans's "analogy" (*Ibid.* p. 137) and Quinlan's "semantic memory program" (*Ibid.* p. 142). His analysis of the problem with this form of AI is that all these are attempts were very specific, and no attempt was made to solve the underlying issues of semantics, which humans obviously have a natural solution for.

2.3.2 Part II – Assumptions Underlying Persistent Optimism

In part II Dreyfus turns to an analysis of the assumptions behind both forms of AI. He details four assumptions that are in the heart of AI as a research programme, at least one of which a researcher must embrace in order to pursue artificial intelligence:

1. The **biological** assumption (*Ibid.* p. 159-162), that at some level humans operate in a digital manner. Every generation thinks in terms of its latest technology (Bolter, 1984), and so we should excuse Aristotle of thinking of the brain as a cooling device, and we should therefore forgive those alive in the later part of the 20th century for thinking in terms of computers. But even if the brain *were* some sort of computer, there is no evidence that it would be in any way similar to our computers. All evidence we have points to the brain *not* being digital (see also (B. C. Smith, 2005)).

2. The **psychological** assumption (Dreyfus, 1979, pp. 163–205), that there is some *mental* level which is digital (this is at the heart of cognitive science, see (Boden, 2008)). Dreyfus doubts that there is any non-metaphorical way in which we can discuss information-processing notions like "list processing" as anything but one of the things that the mind is *capable* of doing – but the psychological assumption and the cognitive research paradigm require that the mind be *constituted* in these discrete, digital information-processing terms.

3. The **epistemological** assumption (Dreyfus, 1979, pp. 189–205) is that perhaps neither the brain nor the mind are digital, but just as the planets are not actually calculating their trajectories around the solar system, nonetheless their trajectory *can* be calculated. In a sense this is the assumption of AI-as-technology as opposed to cognitive or brain simulators. In AI-as-technology all we need is to simulate the *behaviour*, not the precise actual mechanism that produced it. The optimism of this programme is to a large part based on the success of physics and the subsequent technology. Dreyfus doubts that any non-arbitrary human behaviour can be formalised.

4. The **ontological** assumption (*Ibid.* pp.206-224) is that all facts are enumerable, and can be presented to a computer. The idea that all facts can be made explicit has a history going as far back as Plato, but received its most recent prominent statement in the Tractatus: "*The world is the totality of facts, not of things*" (Wittgenstein, 2001b, para. 1.1). Minsky attacks this notion with his estimate that "sensible behaviour" would require between 10^5 and 10^7 facts. This gives rise to the "large database problem", and the related frame problem. These

> problems arise, for Dreyfus, not from human intelligence but from the ontological assumption, which is only one possible interpretation of human intelligence. His alternative is a flexible intelligence that is always already in a situation, and therefore has no frame problem nor has any need to look up relevant facts in a large database.

Dreyfus turns (1979, pp. 231–234) to giving his alternative by pointing out that the Platonic-logical tradition and the computer industry are forces so strong that they overwhelm all before them, but one must at least try to be aware that the direction they are taking human culture is not the only possible one, and the assumptions involved (above) are not axioms we should never question. Presenting an alternative as a *scientific* theory would be falling back into this Platonic tradition, because a scientific explanation, these days, *requires* separating any object into atomic parts, and so subsumes it under the very framework Dreyfus is trying to reject. Dreyfus presents as an alternative phenomenology. Phenomenology (to Dreyfus) is diametrically opposed to a mechanical explanation. Its explanations aim to find the necessary and sufficient elements that go into human behaviour. Dreyfus draws mainly on the work of Martin Heidegger and Maurice Merleau-Ponty.

Some points of nomenclature are due. In everyday speech the terms world and universe are used near-interchangeably. In the parlance of phenomenology the term “universe” means the objective perspective, planets and other things “out there”, while the word “world” means something more akin to “the world of the child” or “the world of the 17th-century London carpenter” - it is the subjective world of a particular individual. This is related to the Kantian distinction between phenomena and noumena. Likewise, a “situation” is conceived as pertaining to a particular subject, while a “state” is physical, external, and objective. So a state is in the universe, and a situation is in a world. Similarly “coping” in a situation is the phenomenological term for “responding” to a state, being “situated” is the subjective side of being in a specific place and time, etc.

2.3.3 Part III – Alternatives to the Traditional Assumptions

Admitting that his account is less precise, Dreyfus examines three areas:

1. The **role of the body** in intelligent behaviour (*Ibid.* pp. 235-255)

   Descartes made an argument that machines can only be in a small number of states, and therefore cannot respond to all the complexities of the world (without an immaterial soul). Convenient as this would be for Dreyfus, he must admit that the number of possible states of a modern computer is so vast that this argument no longer holds. Dreyfus argues that what is missing from a machine is not a soul but being "*an involved, situated, material body*". AI has done well in all the "higher" parts of the mind, like logic etc., but has failed miserably in the parts of behaviour that we share with animals. Our perception is global (Gestalt), with the overall meaning determining the parts of speech and the phonemes. Perception involves a distinction between the foreground and background, with most of the retina-input ignored as background. Our perception involves an "outer horizon" - the limits of what we notice. A computer, by contrast, has to either process some data explicitly or not at all. Perception also involves an "inner horizon" of perceiving objects as being whole, even when we only see the upper side of a table and three of its legs, we perceive it as *having* an underside and four legs.

   Skills need to be acquired, and perception is also a skill, no less in vision or feeling than in language understanding. Unlike computers, in humans there is a clumsy rule-following phase later replaced by a smoother skill (see the video examples in section 7.5.4), or a gestalt of skills. This requires practise, and practice requires being involved with a body. The body is already skilled in acquiring skills. Skills are acquired by what Merleau-Ponty calls "maximum grasp", which is a continuous monitoring of the situation while measuring how well one is coping with the situation (*Ibid.* p. 250). This is situation and goal dependent. Science may require a detailed description of every motion, but our skills as such do not – birds are not aeronautic engineers – they just fly. Skill also allows us to adopt tools in a sense as part of our own body, as a skilled carpenter adopts his hammer, or as we all do with our main language.

2. Orderly behaviour **without recourse to rules** (*Ibid.* pp. 256-271)

   Our (rationalist) philosophical tradition believes that every orderly behaviour

can be formalised in terms of rules. In open-structure problems, there are (at least) three phases: one needs to find out which elements can possibly be relevant, Which actually *are* relevant, and which of these are essential. All these distinctions vary with the situation.

Dreyfus's alternative view is that the situation is *not* a collection of data that needs to be sorted into relevant/irrelevant categories etc., but is always already imbued with meaning. A punter deciding on the horses does not use a database of all facts about horses and people, but always already knows what it would be like to be upset by a parent's death or to be disappointed in love, and knows that such events would effect a jockey's performance not because of some arms-length analysis of an alien situation but because the punter is involved in the same field of concerns which constitutes being human, so he always already knows "what it would be like" for a jockey to lose his mother as opposed to losing his watch.

The human world has a unity to it, that arises out of it being *my* world, with *my* concerns, and with implements being there for purposes, *my* purposes. The AI universe of disjoint pixel inputs being modelled into a "universe model" of objective objects is devoid of all meanings, significance, or unity. "*nowhere* [in AI] *do we find the familiar world of implements organised in terms of purposes*" (*Ibid.* p. 267).

Dreyfus continues with the discussion quoted above (section 2.1), and summarises that "to avoid inventing problems and mysteries we must leave the physical world to the physicists and neurophysiologists, and return to the description of the human world which we immediately perceive" (*Ibid*. p. 271).

3. The **situation as a function of human needs** (*Ibid.* pp. 272-280)

   People's needs cannot be pre-specified. One cannot say there is a need for "food", "shelter" etc., and thereby predict human behaviour except in the most cruel and extreme situations. People discover what they want, what they need, as an act of creative self-discovery. One can say that a man needs love, but that need is never fully specified as just "love". After that man falls in love with a specific woman, his need is for *her*, not for "love" or "a woman" in general. We

explore the world and stumble upon things that we later say that we always needed – and these are always specific and arrive in the day-to-day involvement in the world. So no pre-specified means-ends-analysis, with a predefined set of "ends" will come close to implementing the way human intelligence operates.

Moreover, not only the motivations and needs of a human are not pre-determined, even the correct description of a situation, even for a committed scientist, is not pre-fixed but is determined by the observer, subject to her pre-conceptions, or "paradigm" to use Kuhn's terminology (see also section 3.3.3.4). Man's nature is indeed so malleable that Dreyfus worries (*Ibid.*, p. 280) that if the current infatuation with computers and the AI way of thinking continues then humans will think of themselves more and more in terms of the pre-specified means-ends paradigm, and the danger to humanity will not be from super-intelligent machines but from sub-intelligent humans[3].

2.3.4 Dreyfus's current position

Dreyfus (2007)[4] uses the frame problem as key to understanding how different researchers have tried to achieve AI. He takes it for granted that the only AI worth having would be "Heideggerian AI". He quotes approvingly from Brooks, Agre, Wheeler, and Freeman.

Brooks is quoted approvingly for objecting to representations, and looking for a non-GOFAI way of making robots, but his "r*obots respond only to fixed features of the environment, not to context or changing significance. They are like ants*", he also points out that Brooks's systems do not learn.

Dreyfus calls Agre a "pragmatist". Agre (with Chapman) is explicitly Heideggerian, but objectifies the Heideggerian readiness-at-hand, and programs none of the *experience* of skilful coping. Agre was "*putting his virtual agent in a virtual world where all possible relevance is determined beforehand*", and therefore cannot account for learning, or new relevancies.

Moreover,

---

3 Some would say that the wholesale adoption of American business-school methods of management in areas other than business is already doing this (Thanks to Blay Whitby for pointing this out).

4 The self-same talk has been repeated on other occasions, and appears with later dates.

> "*Agre's Heideggerian AI did not try to program this experiential aspect of* ***being drawn in by a solicitation***. *Rather, with his deictic representations, Agre objectified both the functions and their situational relevance for the agent. In Pengi, when a virtual ice cube defined by its function is close to the virtual player, a rule dictates a response, e.g. kick it. No skill is involved and no learning takes place*" (emphasis added).

Dreyfus is perhaps most positive about Wheeler's work - He "*agree*[s] *it is time for a positive account of Heideggerian AI and of an underlying Heideggerian neuroscience*". However he objects to Wheeler's reintroduction of representations, and to Wheeler's adoption of Classic AI's concept that humans are involved with *problem solving*.

Dreyfus suggests that recent research into neurodynamics in rabbits (by Walter Freeman) provides a promising start to understanding cognition correctly. However, this effort is still quite far from producing any usable technology.

Dreyfus maintains that "*most basically we are absorbed copers*", not problem-solvers, or cognitive agents, or planners, etc. He continues that "*at its best, coping does not involve representations or problem solving at all*".

For more about phenomenology see section 3.3.1.5.

2.4 Winograd & Flores

Winograd and Flores pick up the project of critically thinking about AI, and take Dreyfus as one of several sources for their discussion. They describe an intellectual current central to western civilisation (and to Dreyfus's critique) they call "the rationalistic tradition" (Winograd & Flores, 1986, pp. 14–26) It is characterised by approaching any and all problems in a series of steps:

1. Characterise the situation in terms of identifiable objects with well-defined properties.
2. Find general rules that apply to situations in term of those objects and properties.
3. Apply the rules logically to the situation of concern, drawing conclusions about what should be done.

In most of the analytic or English-speaking tradition of philosophy questions about points 1-2 are often neglected as being antithetical to the scientific project, which is precisely about explaining and making predictions about situations in clear-and-distinct

terms, using general rules. Most attention in English-speaking philosophy, (especially as “the hand maiden of science”) goes to finding better ways of applying 3. This rationalistic approach is at the heart of science, and enjoys all the prestige that comes from science's success. For many this is the right (or possibly the only) way to think. Anyone objecting to this way of thought is accused of “having religion up their sleeve” (Boden, 2008, pp. 251, 324) or of mysticism, or of being incomprehensible (McCorduck, 2004, p. 230).

The rationalistic orientation pervades, therefore, not only computer-science and AI, but also scientific psychology, management theory, linguistics, and cognitive science (not least thanks to Simon and other scholars being active in several of these fields simultaneously). Careful thinkers within this tradition do admit to its limitations, but in daily work from computer-science through cognitive science to psychology and the social sciences, far too often this methodology is taken for granted, not only in what answers are accepted, but also in what questions are even allowed.

Winograd & Flores discuss three alternatives to the rationalistic view:

2.4.1 Understanding and being

Like Dreyfus (section 2.3) Winograd & Flores introduce Heidegger (Winograd & Flores, 1986, pp. 27–37) as a major possible source of ideas for cognitive science and AI. However, they expand on Dreyfus by going into Hermeneutics. **Hermeneutics** (the theory of interpretation) began as a theory of the interpretation of texts, especially religious texts, such as the bible. One of the important insights of phenomenology is that people, when interpreting art, music and texts use similar approaches to when interpreting anything else, e.g. making sense of situations they are in. In a sense, studying the practices people apply to understanding art or texts can be used as a test case for developing sensibilities for understanding how people make sense of the world in general.

One of the key observations of hermeneutics is the **hermeneutic circle**. The idea is that the whole is always understood in terms of the parts, and the parts can only be understood as part of a whole. So how can we understand *anything*? Different thinkers address this problem. This is somewhat analogous to the frame problem in AI – the problem of finding the context in which to understand some input.

One of Heidegger's key points is denial of the dualism of subject and object. We never have an experience of subjectivity without it being directed towards an object, nor of an object unobserved by a subject. What *is* going on (by Heidegger) is a unity he calls being-in-the-world, the **ongoing encounter** between the human (dasein in his parlance) and his world. That is not to deny that there may be an objective universe out there – Heidegger is just saying that is not what is going of *phenomenologically* – phenomenologically we *are* "being in the world" - an ongoing encounter between "**dasein**" (approximately, a fancy term for human) and dasein's world. This encounter is not neutral, disinterested, platonic, scientific, but caring – the world in encountered inside a context of dasein's concerns.

The subjective and the objective cannot exist independent of each other. They are theoretical (non-existent) polar opposites of what is really going on – a process of encounter, which is identical to the process of interpretation (of the world and "objects" by dasein). "*The interpreted and the interpreter do not exist independently: existence **is** interpretation, and interpretation is existence*" (*Ibid.* p. 31).

Hans-Georg Gadamer (b 1900, d 2002) continued Heidegger's work in the field of hermeneutics, with Heidegger's approval (Malpas, 2013). Two of Gadamer's key concepts are tradition and prejudice. He shows that all thinking is done within a context of a tradition (if only minimally in that no one person invented language). Plato was operating in an intellectual tradition that already included Homer, Pythagoras and Socrates, just as present physicists operate in a tradition that includes Occam, Leibnitz, Newton, Einstein, Feynman, etc.

Gadamer re-examined **prejudice**, showing that prejudice does not only have the negative meaning it has in our daily parlance, but is also a necessary condition of any understanding. Total "openness" (let's take that as the theoretical opposite of prejudice) cannot understand anything, because it does not have any categories or language with which to interpret and understand anything. Physicists *need* their commitment to mathematics to gather and *measure* their observations. A businessman *needs* the notions of value and money in order to examine opportunities. An interesting example of prejudice in action in this positive sense is that our current society (following centuries of struggle against dogma, racism and slavery) has a very strong prejudice against the

notion of prejudice itself.

A few points about Heidegger's position:

Our implicit beliefs and assumptions cannot all be made explicit. There is no neutral position we can occupy where we examine our own thought processes without these processes being active. The inevitability of this circularity, and hence the fact that the project of human self-exploration is endless – these should not discourage us, as in every iteration we can find out more about how we operate, only this information will never be clear and explicit, complete and objective. Being human is not an engineering project.

Practical understanding is more fundamental that detached theoretical understanding. Our mind is not designed/evolved to deal with platonic theories. Understanding the nature of our mind, which is the same as understanding the nature of interpretation, should start and end with our encounter with everyday life, not with advanced academic matters, which are highly contrived (e.g. formal grammar, mathematics).

We do not relate to things primarily through having representations of them. What we have, in our practical and day-to-day mind, is familiarity and skills of dealing with situations and the world – not some engineer's schema of the objects around us (this is the main thrust of Dreyfus's attack on AI in (Dreyfus & Dreyfus, 1986)).

Meaning is fundamentally social and cannot be reduced to a meaning-giving activity of individual subjects. This point is similar to Wittgenstein's point about meaning being given within the context of a language-game, where there are socially-agreed rules (e.g. "anthropology" is not a colour).

Thrownness is one of Heidegger's basic concepts. It describes the human predicament, that life itself, and everything and every moment in life always starts before one is quite ready. We are never fully prepared for anything, nor do we fully understand anything. Winograd & Flores demonstrate thrownness with an example of having to chair an important meeting:

- You cannot avoid acting
- You cannot step back and reflect on your actions

- The effects of your actions cannot be predicted
- You do not have a stable representation on the situation
- Every representation is an interpretation
- Language *is* action

This may be more clear and acute in chairing a meeting, but it is true of every waking moment.

Perhaps the most central Heideggerian concept is "**readiness to hand**". A hammer (the classic example) is not given to us in terms of its objective properties, nor is it really present for us when it is in its correct context and in use. A Hammer shows up to us with all its technical detail when it breaks, or when it is completely in a wrong context, like in our salad, or displayed in a museum. When a workman is driving a nail, the hammer is hardly more present to him than his hand, or the tendon *inside* his hand. So equipment, in its correct context, is transparent (Heidegger calls this "ready to hand"). The same equipment becomes present (and an annoyance) when the hammer breaks, or isn't where it should be, or is in a wrong context.

2.4.2 Cognition as a biological phenomenon

In nature, any system that has a boundary and tries to control the environment (outside) for its benefit using its body (inside) can be considered a form of life. Forms of life include bacteria, horses, people, and cities. Any attempt to look at people as a form of life would be biological, and would be based on physics and chemistry. Zoology, anatomy and evolution are a "macro" view of this level of description.

In AI, examples of this approach are neural nets, a-life and genetic algorithms.

In philosophy of mind, this approach is exemplified by the theory of autopoesis by Maturana & Varela. Winograd & Flores (1986, pp. 38–53) discuss this level as "Cognition as a biological phenomenon".

2.4.3 Language as listening and commitment

An important aspect of our culture is the way that promises, commitments, etc. are used in society, and specifically in organisations. These are of paramount importance to Winograd & Flores (pp. 54-69), and they propose software arrangements to facilitate

such social structures.

Oddly, after all the detailed assessment of “computers and cognition” the book takes a sharp turn away from AI and goes on to exalt the virtues as software for managing group-work situations. This was later implemented and called “group-ware”, the most widespread example was “Lotus Notes”, now renamed “IBM Notes” (IBM, 2014).

2.5 Hermeneutics and Gadamer

Winograd & Flores introduce Hermeneutics and Gadamer only briefly (see section 2.4.1), but a bit more is necessary for this thesis.

This is a (necessarily) schematic introduction to hermeneutics and how it relates to phenomenology. Initially I will treat these as entirely separate things. In the 20th century these merged to a degree, and were also (at times) barely separable from existentialism and literary criticism – especially in the personae of **Martin Heidegger** (b. 1889 d. 1976) and **Jean-Paul Sartre** (b. 1905 d. 1980) (P. Watson, 2001). Luckily, for our purposes here we needn't delve too deeply into this tangle of intellectual traditions, nor do we even need to have a full grasp of Phenomenology or Hermeneutics – a sufficient summary will be provided here.

2.5.1 Phenomenology, hermeneutics, and other disciplines

“***Phenomenology*** *is the study of structures of consciousness as experienced from the first-person point of view*”, and is arguably as old as Buddhism, but (at least in the west) it “*came to full flower in* ***Husserl***” (b. 1859 d. 1939) (D. W. Smith, 2013). Heidegger, Husserl's student, revolutionised the ontology implied by phenomenology – for Husserl and most of his students questions of being or ontology were “bracketed” or set aside, leaving the phenomenologist with essentially an idealist ontology (D. W. Smith, 2013). Heidegger argues that we cannot understand the human condition other than in the human's involvement and interaction with the world (being-in-the-world), and makes this very interaction into his new ontological foundation. In recasting this human condition of interaction as basic (instead of the more traditional idealism or materialism), Heidegger points out that fundamental to a human's interaction with the world is the act of interpretation, of making sense of the situation one is “thrown” into (Heidegger, 1962, pp. H135, H298).

By contrast, **hermeneutics** (the theory of interpretation) was for most of its much longer history not a philosophical tradition but rather the theory of how to correctly understand religious texts (Ramberg & Gjesdal, 2014). Arguably hermeneutics is at least as old as the Pauline epistles in the new testament, however it is with **Martin Luther**'s (b. 1783 d. 1546) "… *'Sola Sscriptura' that we see the dawn of a genuinely modern hermeneutics*" (*Ibid*.). This protestant injunction, that the bible should be interpreted only on its own terms (without any reference to Catholic tradition) is probably the first explicit statement of a *policy* or *principle* by which interpretation of a text should be carried out (*Ibid.*).

Speaking against Cartesian notions of understanding ("clear and distinct"), **Giambattista Vico** (b. 1668, d. 1744) "*argues that thinking is always rooted in a given cultural context. This context is historically developed, and, moreover, intrinsically related to ordinary language*" (Ramberg & Gjesdal, 2014).

The Romantic tradition, captivated as it was with holy texts from varying traditions (P. Watson, 2006), gave rise to the first theory of understanding in general, by **Friedrich Schleiermacher** (b. 1768, d. 1834). He discussed the alien nature of foreign texts, and called for particular attention to our prejudices, so we can understand texts under their own alien context. He does not guarantee that such strict awareness of prejudice and openness will lead to a correct understanding of a text (that may be impossible). However such openness is *necessary* for understanding, and is required not only for foreign texts but for any type of communications. Because neither is such an openness ever complete, nor is our information about the context of the writing of the text full, no interpretation is ever final. Schleiermacher's work was seen as the beginning of a "*critique, in the Kantian meaning of the term, of historical reason*" (Ramberg & Gjesdal, 2014).

The next major thinker in hermeneutics was **Wilhelm Dilthey** (b. 1833, d. 1911). He distinguished "living experience" which is how each of us experience ourselves, from "understanding" which is how we more systematically understand the world outside us and others. He claimed that true self-awareness can only be achieved when one understands oneself in the same terms one understands others. In understanding history and historical texts one should combine (what we would now call) empathy, i.e. a

“living experience” identification with the historical characters, with “understanding”, which is a more rigorous “from the outside” observations. The “living experience” component allows the historian to form hypotheses about history, while the “understanding” part allows one to critique such thoughts, and see how well they stand to reason (Ramberg & Gjesdal, 2014). This contrast between a creative and critical phase in intellectual work can be seen as a precursor of the context of discovery / context of justification distinction (see section 4.2.4).

2.5.2 The hermeneutics of Heidegger and Gadamer

For Heidegger interpretation is not only a matter of understanding texts, but of our entire mode of being, which is continuously involved with comprehending the world and acting in it – hence hermeneutics becomes one and the same project as phenomenology, and this joint project becomes the new ontology (Ramberg & Gjesdal, 2014; Winograd & Flores, 1986, p. 31). Heidegger was concerned with many issues in phenomenology, and viewed the specifics of hermeneutics *as such* as a sub-field, the detailed exploration of which he later entrusted to a large degree to Gadamer (Malpas, 2013, Chapter 4).

Gadamer viewed hermeneutics not only as the theory of understanding ancient texts and art in general (Gadamer, 2004, pt. 1) but also, and perhaps mainly, as the act of continuously understanding/interpreting all situations. In this sense, interpretation is an unceasing activity (during at least most of waking hours).

Here is an example (my own) of what is meant by interpretation in this context. Consider the following:

- הכלב מכוער
- Ha-kelev meh'oar
- Il cane é brutto
- The canine is brutish
- The dog is ugly

At this point you may be perplexed by this strange list, as one would be with any other strange sequence that is presented with little warning. In a sense I just caused you to be

“thrown” onto this unusual list, and to the urgency of making sense of the situation, but help is at hand... The lines all convey the same meaning (in different alphabets, languages and dialects). Note how much easier it is to interpret (for an English monoglot) these examples the further down one goes. Note also that as an English-speaker you may be further interpreting the situation and objecting that “brutish” does not mean the same as “ugly”, but you also may be aware that in the Italian “brutto” does actually mean ugly, and may further be aware of how such words change meanings over the centuries and the geographic distances involved. All these thoughts are interpretative – they are attempts to make sense of a situation, at this instance the situation at hand is the bizarre list above. *This* interpretative effort is what is meant when Heidegger, Gadamer and others say that interpretation is our “mode of being” or suchlike expressions.

Interpretation (in the sense that interests us here) is the ability to “follow along”, to “make sense” of the “inputs”. In following along with (say) a song, this is easier with a familiar tune than it is with foreign music. The crux (here) of the knowledge or skill accumulated as we become more familiar with a situation does *not* consist of beliefs - we have no position on the ugliness or beauty of a dog we have never seen. What *is* being formed is an *interpretation*, an understanding, a grasp – before (and not requiring) any judgement.

Gadamer's view of interpretation is contrasted with the objectivist school that viewed the purpose of interpretation as reaching the “true” or “objective” meaning of the text (Winograd & Flores, 1986, p. 28). Gadamer views the process of interpretation as a meeting, or a clash, or a merger, of two “horizons”:

1. the brute facts of the text (which word is where *in the text*) and
2. the sum-total of all knowledge, attitudes and prejudices of the reader.

Hence the name of Gadamer's magnum opus (2004), “Truth and Method” - Truth stands for the brute facts of the text, and method is all the wisdom the reader brings to bear.

The word “prejudice” has a chequered history – in earlier hermeneutics, and also in common modern usage, this term is seen as negative, and indeed one of our culture's strongest prejudices is a prejudice against “prejudice” itself. However this is (by

Gadamer) a very narrow reading of the term: prejudices are unavoidable. We only read texts that are available to us by some accident of history, and once some knowledge is acquired, it will colour our understanding of related topics for the remainder of our life. So in a sense everything that we bring to the process of understanding, all our history, everything that falls under “method” - these could all be called prejudices, in that they colour how we will see all things (Gadamer, 2004, pp. 267–304).

Heidegger has already pointed out that in the process of interpretation we encounter the “hermeneutic circle”: We understand the whole in terms of the parts, but we also only understand the parts in terms of the whole. Gadamer adds another view of the hermeneutic circle: The meaning of a text is determined (at least in part) by the “method” or “prejudice” (or “mindset”) of the interpreter – and also the entire cultural being of the individual is constituted in the various influences on herself, including the very text under study. So the text (part) determines the reader, who in turn part-determines the meaning of the text (Winograd & Flores, 1986, p. 30).

How this work relates to a Gadamerian understanding is discussed in section 8.4.2. One should recall that Gadamer was exploring a specific aspect of Heidegger’s world-view – namely the hermeneutic aspect. In exploring Gadamerian AI we are exploring one aspect of what Heideggerian AI would be. This is in line with (Dreyfus, 2007)’s call for a more Heideggerian AI, but in saying that it is “a step in the right direction” one is still threatened by the “first step fallacy” (Dreyfus, 2012)

2.6 Literature review: summary

Dreyfus blames the overwhelming success of physics for the biases of the AI research programme, though he does not give this way of thinking a consistent name: “Platonist”, “intellectualist”, “mechanist” and “they” (sic) figure often (Dreyfus, 1979, pp. 191–202).

Winograd & Flores (1986, Chapter 2) name this “the rationalistic tradition”, and characterise it as aiming to solve problems by using three steps:

1. Characterise the situation in terms of identifiable objects with well-defined properties.
2. Find general rules that apply to situations in term of those objects and properties.

3. *Apply the rules logically to the situation of concern, drawing conclusion about what should be done.*" (Winograd & Flores, 1986, pp. 14–15)

This critique of science-like thinking extending beyond the bounds of its competence is widespread in the humanities and social sciences, where it is called scientism (the term is derogatory) (Bannister, 1991). Usually the people using the word "scientism" in social science are fighting against what they see as an overextension of scientific practise to cover areas that should be given a more qualitative or a more nuanced treatment.

In doing a literature review, one is supposed to lay out everything which is relevant to the subsequent discussion. This is premised on an agreed delineation of disciplines, and on the convenient fiction that areas of thought can be delineated. This may be practical in some areas, especially in science – but it is not the case here. Our field (philosophy of artificial intelligence), perhaps even more than others, is populated by people whose ways of thinking are products of a certain history.

-

To summarize the introductory chapters: There is a dearth of fundamentally new ideas in AI. Two under-exploited sources of ideas near the existing literature were found: The main one is the debate about phenomenology, and another one is Papert's admission of a mass-pretence about thinking and logic (quoted in section 1.1). There are always reasons why some areas of research are left neglected, and these should be occasionally reviewed. In our case the main barrier to overcome is scientific overreach, in two senses: One is that researchers have used science-level requirements in a technological field, unnecessarily being too stringent. The other is that categorical methodological judgements that make sense in science were applied as universal truths also in AI as a technology. First and foremost amongst these is the judgement against introspection, formulated most centrally by J.B.Watson (1913), but carried forward forcefully by Simon (see section 2.2) and spread by his towering presence throughout the AI community. The main tool for combating such prejudices will be distinguishing different perspectives and different types of truth requirement, e.g. between science and technology (see section 1.3.2).

3 Thesis outline and terms

Table of Contents

3 Thesis outline and terms....49
3.1 Terms of this thesis: “is recommended for developing”....50
3.1.1 “Recommended”....50
3.1.2 “For“....52
3.1.3 “Developing”....53
3.2 Terms of this thesis: “anthropic”....55
3.2.1 Human vs ideal/rational....56
3.2.2 Motivations for human-like AI....57
3.2.2.1 Rational AI's interaction is “clunky”....57
3.2.2.2 The versatility of human Intelligence....58
3.2.2.3 Getting along with people....59
3.2.3 Characteristics of human-like AI....60
3.2.4 Human-like vs anthropic....61
3.2.5 Perspectives and levels in human modelling....63
3.2.5.1 Are there really levels or layers in the mind/brain?....63
3.2.5.2 Multiple levels of discussion....64
3.2.5.3 The cognitive level is problematic....67
3.2.5.4 Simultaneous multiple levels in computers....68
3.2.6 Anthropic AI so far....69
3.2.7 Knowing that vs knowing how, and a hint on data structure....71
3.2.8 Metaphysical non-problems....74
3.2.9 Ethics....75
3.2.10 Anthropic AI: summary....76
3.3 Terms of this thesis: “introspection”....76
3.3.1 Studying subjectivity....77
3.3.1.1 Why subjectivity?....77
3.3.1.2 Locating subjectivity....78
3.3.1.3 What is subjectivity....79
3.3.1.4 Subjectivity can be studied....80
3.3.1.5 Phenomenology, hetero-phenomenology....81
3.3.2 Defining introspection....82
3.3.3 A boundary between introspection and science collapses....84
3.3.3.1 “Thinking aloud” (TA) can be seen as introspective....84
3.3.3.2 Two distinctions between TA and introspection....86
3.3.3.3 Inferences and confusion....89
3.3.3.4 Non-inferential observation is impossible....89

3.3.3.5 A boundary between introspection and science collapses: conclusion. 91
3.3.4 What kind of introspection is recommended.....91

This chapter presents my thesis, “Introspection is recommended for developing Anthropic AI” outlining how some of the main elements will be dealt with in subsequent chapters, explaining all the main terms, and making some of the secondary arguments. This chapter includes the full argument with a finite list of holes, which will be filled by subsequent chapters. These coming chapters will also add details and pragmatics. Later still I will give some working examples and discuss some consequences of this project.

There are four key terms that will be used, that are worth previewing here as a preliminary outline.

1. This thesis is about **human-like AI**, as opposed to rational/idealised AI (see section 3.2.1).
2. Within human-like AI, My focus is on **Anthropic AI**, approximating the underlying mechanisms of humans-as-such, rather than the accomplishments of western, modern, well-trained adult people (see section 3.2.4).
3. **Subjective methods** in AI have been relatively neglected though they give us some access to how we work, at a level that is practical to simulate, rather than (say) simulating every cell in a brain (see section 3.3.1).
4. Introspection is how we can access subjectivity (see sections 3.3.2-3.3.4).

This chapter will start with the middle terms “is recommended for developing”, and will proceed to discussing the purpose, “anthropic AI” and the means, “introspection”.

3.1 Terms of this thesis: “is recommended for developing”

This thesis's main claim is that “introspection is recommended for anthropic AI”. Let's look at the middle terms: “recommended for developing”.

3.1.1 “Recommended”

Recommending something is not a guarantee that it would always work. It is an

assurance that one has reason to believe that it would work (or be profitable) in enough of the cases to make it worth pursuing. In other words though a recommendation is not a guarantee, it is also not vacuous.

Here is a summary of how this recommendation will backed in detail in subsequent chapters: After introducing the terms in this chapter, in chapter 4 I will show that using introspection for AI development is *permissible* (even though it was treated so far in the literature as illegitimate), and in chapter 5 I will show that using introspection is a plausible way of gaining access to a description that in many cases suffices for the reproduction of human skills.

As detailed in chapter 4, introspection was forbidden by Watson, as of (1913) (see sections 4.2.1). Watson forbade introspection for psychology, as he wanted to strengthen psychology's claim to being a science, and to make research into human behaviour contiguous with research in animal behaviour. This was motivated both by the prestige of the “hard” sciences, and by a Darwinian effort to eliminate any special status of humans over the rest of the animal kingdom (Costall, 2004, 2006). Regardless of much debate about introspection in the last 100 years (6 different positions on how introspection relates to AI will be enumerated in chapter 4), not a single AI developer embraces introspection wholeheartedly and uses it to build working systems. This is shown to be misguided since the type of truth required in technology development is quite different from the one required in science. Additionally (even is we were to assume a scientific discourse) the attitude of AI developers seems to ignore the distinction between the context of discovery and the context of justification. The conclusion of chapter 4 is that introspection (even the worst type, discussed below in section 3.3.3) is an *acceptable* basis from which to build AI. Even in cases in the past where introspection *was* partially used as such a basis, that usage was done timidly and apologetically, as if “in sin”.

As will be shown in detail in chapter 5, introspection is the basis of most attempts to turn mental skills (knowledge how) into any sort of communicable form. Since some skills are transmitted in human culture for many thousands of years, the very survival of civilizations for more than one generation is a living testimony to the (at least frequent) success of introspection. Introspection and communication succeed in capturing enough

of the essence of skills so that they can be communicated from one generation to the next. An argument can be made that the very (evolutionary) reason we have consciousness is to allow the communication of acquired skills from one individual to the next, including the young. This can lead to the interesting aside of asking who is evolving, humans or civilisations, and who owns whom - do humans have cultures, or do cultures possess humans? Conveniently, this is outside the scope of this thesis. The conclusion of chapter 5 is that introspection is a *plausible source of ideas* for anthropic AI.

One can recommend various processes for AI developers. One could recommend reading poetry, meditating, taking a walk or sitting on a comfortable chair. One could even produce evidence that some of these recommendations do improve AI research. Here my recommendation is based on a more intrinsic link between introspection, skills, and anthropic AI. Having shown that introspection is both acceptable and plausible, details of my precise recommendations are found in chapter 6 and examples are provided in chapter 7.

One must bear in mind that at the moment no AI researcher is wholeheartedly embracing introspection *and* writing code. This is the crux of this thesis – **introspection is recommended for developing anthropic AI**.

3.1.2 “For“

This thesis recommends introspection for developing anthropic AI. So we are looking to make anthropic AI *based on* some introspection. The exact nature of this “based on” relationship and many examples of different types of AI and what these are based on will be found in section 6.1. For now suffice it to say that Y, a design for an AI system, is based on an observation X (that could be an introspective observation) iff:

A) There is a causal link from X to Y.

B) X is the dominant influence on the workings of Y, i.e. there is no significant pollution by some other factor such as a prior theoretical commitment. In our case of AI based on introspection, this would require acceptance of introspection (X) as an acceptable source of ideas, not to be obfuscated or denied; minimisation of attachment to or influence of any theoretical framework, such as

mathematics, logic, or some theory in cognition, psychology, religion, or even phenomenological literature.

C) Corresponding functions are achieved in similar ways (data flows, data structures, temporal order, etc.).

A more detailed version of this definition and how it related specifically to introspection is given in section 6.1.1. Examples of similarities of process and data flow are given in sections 6.1.2 and 6.1.3. For the detailed process of how introspection turns into code see section 6.2.

3.1.3 "Developing"

This thesis concerns development of anthropic AI, specifically the "discovery" or "idea" phase, as opposed to the software development phase. In talking about the processes that go into developing AI, one would benefit from keeping a clear notion of five different minds (or "part-minds") that may be involved in the process, and have different perspectives and concerns. Consider Illustration 3.1 on page 54:

- The **Basis** – This is the base idea, or the inspiration used to build the AI design. It need not be a complete mind, but probably has to be some kind of information-processing or "intentional" entity. Illustration 3.1 presents the examples of logic, mathematics, neural nets, honey bees, introspection, and externally observed behaviour.
- The **AI** program, the machine or robot being built.
- The **Practitioner** who uses the basis as a guide or model and builds the AI system. Examples (in the picture) are Trenchard More, John McCarthy, Marvin Minsky, Oliver Selfridge, and Ray Solomonoff (Knapp, 2008)
- The **Observer**, who may comment on the AI or the process of its development, but is not directly and actively engaged. Examples include Dreyfus, McCorduck, Flores, etc.

- The **King**, or administrator, or the research funding agency. This is “he who pays the piper” or “she who exerts control”, explicitly or implicitly. A prime example from AI would be DARPA[5].

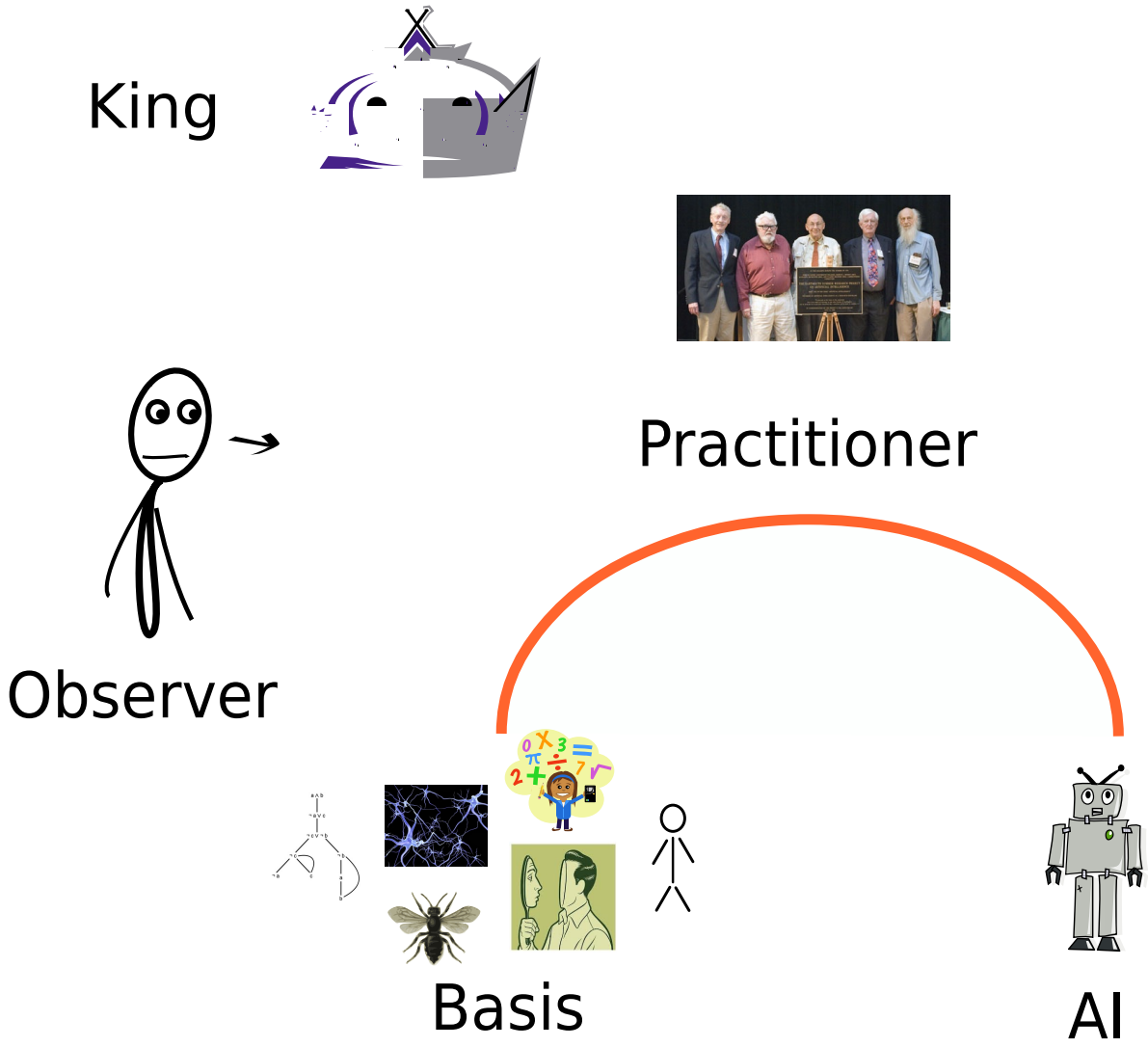


Illustration 3.1: Roles in AI development

In some cases one and the same person can fulfil more than one role, for example Minsky (1991) was both a practitioner and an observer.

This thesis, mainly from the position of an **observer** (bold terms refer to Illustration 3.1), will argue for the **practitioner** using introspection, which means taking both the

5 The degree that AI specifically and computers in general were developed as a military tools is an under-appreciated question (Edwards, 1997). The evidence in the terminology such as “commands” for instructions, etc., is suggestive. This point was made to me by Blay Whitby.

role of **practitioner** and of **basis**. This thesis is *not* about building **AI** systems that will introspect themselves. It is about the AI **practitioner** using his own introspection as a **basis** for his designs. Schematically, this thesis is a plea by an **observer** to the practitioners to use their own subjectivity as the **basis**, the inspiration, for novel **AI**; and also imploring the **kings** to fund such research. In chapter 7 I will take also the position of the **practitioner** in order to provide working **AI** examples of what is recommended here, and in doing so I will also introspect, so I will use my own mind also as the **basis**.

Perhaps it is worth reiterating what was underlined above – this thesis is not directed at introspecting AI systems. This is not because those are a bad idea, quite the contrary. If we want to create truly human-like systems such systems will most probably need the facility to reflect on their own actions, including their metal actions, and hence introspection would most probably be requisite. But at the moment we do not have AI systems of a sophistication where such abilities are a reasonable next step. For the current stage of research, the urgent task is opening up introspection for the practitioner, i.e. recommending the practitioner to use their own subjective experience of their own mind as a basis for AI development.

Some future ideas that would have to be completed before we ask for systems that *do* introspect themselves are discussed in section 8.2.

3.2 Terms of this thesis: "anthropic"

My thesis is "Introspection is recommended for developing anthropic AI". This section is about the aim, the purpose: "anthropic AI". Anthropic AI will be defined as pursuing the computer implementation of (an approximation of) the **base, minimal, human ability** that *allows for* our culture but is independent of it. It will be contrasted with enculturated AI, specifically the prevalent AI that assumes as its target a western, modern, well-trained and adult intelligence.

I will first discuss human-like vs rational/ideal AI, and then will distinguish anthropic AI as a sub-type of human-like AI.

3.2.1 Human vs ideal/rational

Russell and **Norvig**, in their canonical introduction to artificial intelligence (S. Russell & Norvig, 2013, pp. 1–4) introduce a distinction between human-like AI and rational AI. Most of the AI work in the past 60 years has been of the rational, idealised kind. Rational AI aims at correct, or best-possible, solutions. Human-like AI aims at imitating humans, with all their frailty and hopefully ingenuity. Russell and Norvig's distinction is a slight oversimplification. There are other actual existing non-optimal intelligences other than the human case. In cognitive science there are bodies of work on insect and swarm intelligences, and other animals' intelligence. This an active field of research both in industry (Raibert, Blankespoor, Nelson, Playter, & others, 2008) and science (Baddeley, Graham, Husbands, & Philippides, 2012), but these are not directly relevant in the short term to human-Like AI, and will therefore not be further discussed. They also make a distinction between thinking and acting, which is not relevant here.

As a rule of thumb, if an AI system does not **make mistakes** or if you can prove meaningful theorems about it, it is rational AI. There is nothing wrong with rational AI, which is a thriving industry. But shunning the human-like because it is not mathematical enough, not neat enough, or because our scientific methodologies do not apply easily - that would be neglecting an area of research for no better reason than that it is difficult.

It is important to note that not only is human-like AI neglected in research, it is nearly entirely ignored in *teaching* AI, especially in computer-science settings. In (S. Russell & Norvig, 2013, p. 5), after discussing this distinction, the authors declare that their "*text concentrates on general principles of rational agents and on components for constructing them*", and proceed to use the rest of their 1090 pages to teach rational AI alone, possibly misleading undergraduates into the impression that that is the only sort of AI that exists. This is understandable for the authors, as human-like AI is a fringe area, but ignoring it completely is alarming considering that their book is used by 1306 schools in 116 countries (S. Russell & Norvig, 2016).

There is nothing wrong with rational AI. The only problem is that it is (by definition) *not* about humans as they *are*, but insofar as it is about humans at all it is about how they *should be*. This distinction should be kept separate from the distinction between AI motivated by technology and AI motivated by science (section 1.5 provides a two-

dimensional map of AI efforts so far).

There is an interesting parallel to human-like AI in human-like robot (hardware) construction. Nick Hockings (Hockings, Iravani, & Bowen, 2014) aims to build human-like hands, using exact replicas of the human anatomy, down to the level of tendons. At the level of tendons he switches abruptly from being human-like to implementing the tendons using whatever techniques and chemistry are available to build "tendons" as similar as possible to the natural. The idea is to *emulate* the natural, not re-create it, but to do it at a level as low as currently technically possible, feasibly.

In creating humanoid robots we need to emulate the underlying mechanisms, but only insofar as possible. There is a point where in the technological interest one must give up and go to modern plastics and 3-d printing (in robotics) or "just program" (in AI, see sections 3.2.5.4, 6.3.5).

3.2.2 Motivations for human-like AI

3.2.2.1 Rational AI's interaction is "clunky"

Motivations for developing rational or idealised AI are clear: computer technology around us would be impossibly difficult to engineer if it were not deterministic, mathematical, and as fast as possible. Advances in machine learning etc. are ubiquitous and dynamically making tangible contributions to the lives of people (nearly) throughout the planet. Examples (just from smart-phones use) include speech recognition, natural language interfaces, navigation systems that dynamically learn the maps and one-way systems of cities, etc.

Regardless of all these advances, even the most fêted and expensively developed (rational/idealised) technologies today (such as Siri) are clunky, "robotic", etc. The idea, common in Japan (Robertson, 2007), of using robots as companions or care-givers and companions to the elderly is greeted with much scepticism in the rest of the world, because of this clunky behaviour (see section 3.2.2.3).

Other examples of existing AI's failure to capture the human way of doing things are its best-advertised achievements: IBM's "Watson"[6] cannot hold a conversation, and the various attempts at (the broad family of tests called) the Turing test are ultimately

6 Named after the founder of IBM, no relation of the psychologist.

exposed as short-term “bags of tricks” with little expandability (more about bags-of-tricks in section 3.2.8). Other attempts at human-like AI have been purely academic-scientific, in the field of cognitive simulation. In any role where genuine, practical understanding of the human way of doing things is required, robots and AI are so far rightly excluded.

Let's look at a few categories of motivations for human-like AI:

3.2.2.2 The versatility of human Intelligence

In the “Ode to Man” in Antigone, Sophocles (2009) expresses his wonder at the abilities of man:

> *Wonders are many, and none is more wonderful than man; the power that crosses the white sea, driven by the stormy south-wind, making a path under surges that threaten to engulf him….*
>
> *And the light-hearted race of birds, and the tribes of savage beasts, and the sea-brood of the deep, he snares in the meshes of his woven toils, he leads captive, man excellent in wit. ...he tames the horse of shaggy mane... the tireless mountain bull.*
>
> *And speech, and wind-swift thought, and all the moods that mould a state, hath he taught himself; and how to flee the arrows of the frost, when 'tis hard lodging under the clear sky, and the arrows of the rushing rain; yea, he hath resource for all; ...only against death shall he call for aid in vain; but from baffling maladies he hath devised escapes....*

Human intelligence is interesting in its general-purpose nature, its ability to achieve such a diverse range of accomplishments. Human intelligence can learn and act in ill-understood and uncertain circumstances, such as “crossing the white sea”. Any system that tries to be human-like must therefore be a learning system. Humans not only learn all their lives (to varying degrees) but also make mistakes, so aiming at some mathematically correct behaviour may well miss much of what human intelligence is about[7].

Possibly the most powerful motivation for human-like AI rather than rational-idealised AI is the fact that *humans invented rationality*, so we can expect human-like AI to be more flexible and have wider application (even if it may be less reliable and less

7 Much work in rational/idealised AI is done statistically, aiming (for example) to be “probably approximately correct” (S. Russell & Norvig, 2013, p. 725). This is idealised AI since it aims to get the probabilities *right*, and the approximations *optimal*.

optimal). This is of particular interest for the situation of generative AI, where AI will hopefully be employed to develop further AI (van der Zant, Kouw, & Schomaker, 2013).

Formal logic, as we know it today, is specifically a western invention, starting in ancient Greece. Even in the modern west, not everyone thinks logically, and even those who aspire to the feat of thinking logically often fail (Ariely, 2009). None of us were born as logic machines, we were (at best) brought up to be skilled in logic, we were educated into and through logic. So there must be an underlying mechanism that *allows* for the emergence of logic. Some evidence that such a mechanism exists can be found in its malfunction in the case of fallacies. This was explored in detail in the case of algebraic mal-rules (Payne & Squibb, 1990).

3.2.2.3 Getting along with people

Human-like AI would be useful in areas where human-like behaviour would be better than rationally-optimised behaviour – where the very essence of the job is to get along with people, where it is key that the computer be easily understood in human terms, and where it would also be useful for the AI system to understand the human way of doing things (why being human-like is key to understanding humans in discussed in section 3.2.3). Areas of applicability would include:

- Car driving: Driving cars by computer has made great strides in recent years, but still has difficulty with pragmatics (Richtel & Dougherty, 2015). Pragmatics are often culture-dependent. What is done, for example with lanes, is different in different cultures: In Brazil, where politics has driven lanes to be narrower, the idea that a large lorry can take two or even three lanes in a motorway is considered normal. Conversely, in Bangkok lanes will form and dissolve on hard shoulders wherever possible.
- Delivery robots: As online shopping becomes the norm, the desire for fast delivery increases, and a demand is created for faster delivery (Amazon is developing a drone for the extreme version of this problem- delivery in 30 minutes (“Amazon Prime Air,” 2016)). A lightweight robot that can deliver packages to doorways would be invaluable. That would require navigating the addresses, negotiating spaces crowded with pedestrians, and understanding such

human communications as handwritten notes saying “If I am not home please leave packages at flat 6”.

A particularly interesting case is care giving robots. In rapidly ageing societies there is a growing need for carers to keep elders company and to serve as interfaces to digital technology, that in turn can help in physical care (Broekens, Heerink, & Rosendal, 2009). Regardless of ethical issues (Whitby, 2011) (see section 3.2.9) such technologies may become essential especially in societies (Japan stands out) where immigration of human caregivers is not a politically palatable solution (TheEconomist, 2013).

Humans have a tendency to anthropomorphise (treat as human) all entities, for example note Aristotle's idea that heavy objects “want to” go downwards. Regardless of efforts by educators to combat this tendency, humans are still predisposed to think that way. This predisposition is seen even in science where we say things like “the system obeys the laws of physics” or “light follows Maxwell's equations”. There is no obedience or following going on – this is all “in our eyes”, in *our* tendency to attribute human characteristics to inanimate objects.

As robots become more ubiquitous, they will be operated by people with fewer skills. Moreover, in the scenario of care for the elderly, one cannot expect any training given to the patient / operator of a robot to reliably last, due to amplified forgetfulness and confusion in old age – so patients will revert to assuming that the robot is human-like.

In the case of elderly patients cared for by robots this will be much worse (Sharkey & Sharkey, 2011). Patients will assume that robots can apply rules “reasonably” while rational-AI systems have no idea of what “being reasonable” may mean. This can cause patients to trust robots to behave in ways that would be expected of a human, but beyond the robot's preprogrammed ability, or worse – against a clear pre-programmed prohibition. Since this expectation by the patients is unavoidable (and may be life-threatening), as technologists we need to rise to the challenge and make the robots as human-like as possible (perhaps within some hard outer boundaries).

3.2.3 Characteristics of human-like AI

Human-like AI would allow robots to form malleable habits, as opposed to rule-based systems (see section 7.5.4) (like driver-less cars picking up the local driving culture, see

section 3.2.2.3). By thinking and acting in a human-like, culture-adjusted way robots would be better understood by humans. Moreover, once this technology is advanced enough, a truly human-like technology would allow robots to form their own (speculative) understanding of human actors in their environment, to speculate about the human intentions, and "behave considerately" towards these humans.

One of the best ways to understand some entity is to be able to simulate it and see how it would react in different situations. The advantage of human-like AI in understanding humans is therefore analogous to min/max's ability to understand a rival similar to itself – for every system, it is easier to simulate, and to understand, a system similar to itself. Uninitiated humans have no idea of min/max, nor does min/max have a notion of humans – but both know quite well how to deal with another instance of themselves, since they possess the mechanisms to understand, or simulate, their peers. They do not posses the technology to understand a system that thinks differently. As long as the domain of action is a formal domain, like chess, rational/ideal AI has the edge, and wins against humans. Once we move into the human arena, we need human-like AI. This does not preclude integrating various technologies together (Minsky, 1991).

None of this is to say that human-understanding behaviour is *in-principle* impossible for any machine that is based on logical-rational principles (see the discussion of logicism in section 3.2.5.2, and of the "formal sandwich" in section 3.3.1.2). Obviously if we were to ever to program a computer to behave like a human, the computer would still run on silicon chips, which implement a formal system (see sections 3.2.5.4, 6.3.5.3, 6.3.5.4).

3.2.4 Human-like vs anthropic

Let's examine the issue of whether we should want to simulate a fully-fledged or enculturated mind, or whether it would be better to simulate the pre-culture, un-enculturated, naïve mind. If we had a good simulation of a fully-cultured mind it would animate robots that would work in *one* culture. If we go deeper and aim to simulate the un-enculturated mind, we would get the *ability* to get the AI system enculturated into *any* culture. Most of AI so far went for the first, enculturated option, in *anthropic* AI we will aim for the later, un-enculturated option.

So why not use the term "human-like AI" as my target? Because behaving

mathematically or logically is a *possibility* for humans, and would therefore be *part* of a "human-like" concept, and I specifically want to exclude such sophisticated highly-trained thought: Being **western, modern, well-trained, and/or adult** – These may all be desirable qualities in terms of current culture and setting (this is a western, modern, Ph.D. thesis) – but none of these qualities is inherent to being human, and none of us (western, modern, well-trained adults) were born that way. We *learnt* (or *were trained*) to be that way. One could argue that being well-cultured is good, but I argue that simulating our "best practices" is *not* what is needed currently in AI:

- Where "best practices" are clear, we can use normal programming, or rational AI. We already have that.
- Where things are not clear, we need a system that can learn the complexities.
- A system may have a better chance of learning these complexities if it learns them in a way similar to how we learn them, i.e. doing it the way we humans do it, which is more versatile and broader in scope then just "western, modern, well educated adults".

Another important point to note is the question of who decides what precisely "western modern well-trained adult", or "best practices" means. A learning system with fewer pre-judged commitments can better adapt to situations.

Distinguishing a human from their cultural or social context is not simple or easy. Probably this is not even completely possible. But we must try, one of the main questions in this approach is: what is it about *this animal* that allows it to participate in *any* society or culture, let alone our modern one? It is generally accepted that this will include intelligence, or the ability to learn skills and habits.

To get at this level we need to get under, or behind, or around, education or training, as a person's training is a social phenomenon of the specific society in which that person was educated. This attempt to get beyond education is probably never completely possible, but should remain an aspiration. How to do this is the topic of chapter 6.

I propose using the Greek for human, "**Anthropos**", to mean the untrained, basic human. This jives with the way that anthropology studies *all* humans, including the "primitive".

Anthropic AI aims at minimally human intelligence – without presupposing any of our cultural heritage (insofar as possible). Anthropic AI is **defined** as the base, minimal human ability that *allows for* our culture but is independent of it.

3.2.5 Perspectives and levels in human modelling

Next I will survey the different levels we can deal with or model the mind, both for psychological exploration and for AI technology development. But first we must slow down a bit.

3.2.5.1 Are there really levels or layers in the mind/brain?

Caution is advised when discussing layers or levels in the human mind/brain. The idea that things are neatly arranged in layers comes from several sources. In engineering it is useful to think in modules and layers, and in software design not only do we have many layers, we even have a hierarchy (in the types of layers), where some layers are more important, and get to be called "platforms" - like Microsoft's "Windows", the "Java Virtual Machine", "IP" (the inter-networking protocol, as in "TCP/IP") etc. In software the levels are usually very well defined, with the interfaces between layers called "API"s. The idea of the mind being *constructed in layers* is seductive but twice-wrong: The mind is not *constructed*, but is an evolved characteristic of the human animal. We have no evidence that there is anything like *layers* inside the human animal.

It is far more likely that the mind is like the gold-bearing reefs of the Witwatersrand Basin in South Africa (Safonov & Prokof'ev, 2006): Gold settled at the bottom of a primordial lake for millions of years, then the lake-bed (arguably initially a layer) dried, deformed and was partially eroded away. Next most of the gold-bearing deposits were buried deep in the ground. Later a large meteorite hit the ground, tearing and throwing up into the air a 300-km wide part of the earth's crust. As this mass of matter crashed in chaos, some of the gold-bearing formations were exposed, with no particular shape. It so happens that half the world's gold comes from these formations. There are no layers involved, just a lot of history.

Coming back to the human mind, the brain's anatomy (its "architecture") is composed of multiple organs, that evolved in different eras. But that is also misleading, in that the older brain-organs continued to evolve, and so there is no ground to treating them as distinct modules or layers - these organs may be anatomically somewhat distinct, but

they function together. It is doubtful we can even delineate distinct mechanisms - there are no clear-and-distinct “layers” or boundaries to be found, nor is there a reason to believe we may find them in the future. The reason people search for such oversimplified models is that it would be very convenient for our western-modern mode of thinking if we could find such layers. Again, the human situation is more complex than the ideal/rational situation, but neglecting the human just because it is difficult may be a good way to get the first few AI systems going, but eventually we need to tackle the human-as-it-is, rather than the western, modern, formalized way we think that we *ought* to think.

Layers are how *we like* to think about machines, problems, etc. It is something we superimpose on the world in order to make sense of it, like a grid on a map (see the term “adhyasa” in Indian philosophy, section 8.4.5).

3.2.5.2 Multiple levels of discussion

Having defined anthropic AI as the base, minimal, human ability that *allows for* our culture but is independent of it, we can turn to examining what the alternative approaches to human-like AI may be. Being part of the modern west, we cannot avoid some superimposing of distinctions at least in this meta-discussion – let’s just bear in mind that any layers are part of the *analysis* and not of the human. We can identify several modes, or “levels” or ”layers” in which one could observe, discuss and try to simulate humans, their behaviour or their intelligence. Each of these levels takes multiple forms, and can have AI approaches associated with it. I am not making any claim with this list, this is for purposes of clarification only (but see section 3.2.5.3).

1. **Atomic, molecular** (or lower)

   In terms of scientific purity, this may be the best level to simulate anything (that is not so small as to have sub-atomic effects). The problem is that we do not have the data (an atomic-level scan of a human) nor the computational capacity. So this is (currently) infeasible, regardless of the waves of optimism unleashed by the human-genome project in the early 2000s (Bower & Bolouri, 2001).

2. **Cellular** (see Dreyfus's “Biological assumption”, section 2.3)

   In a similar vein, simulating every cell or every neuron is a current goal, for the

decades if not centuries to come (Markram, 2012). For *now* it is infeasible. Moreover, if and when we have a full-brain simulation, there is no guarantee whatsoever that the mind inhabiting such a "brain" will be in any way normal, and/or able or willing to communicate with us.

3. **Bio-functional** (cell assemblies) / Neural Networks

   This level simulates fewer individual neurons or (more accurately) neuron-like abstractions. With this kind of modelling researchers are trying to simulate either small parts of the brain, or entire systems using a (rather strong) assumption that whole cell assemblies behave somewhat like a neuron. Another motivation for this research is exploring what *can* be done with neural-nets. These neural nets are also used in technology, unlike the previous perspectives.

4. **Cognitive-theoretic** (see Dreyfus's "psychological assumption" 2.3)

   Cognitive models (such as SOAR (Laird & Rosenbloom, 1996)) and classic symbolic AI propose a computational model for various faculties that underlie individual human activity (Sun, 2008). These models (when used as a scientific tool rather than for technology) are verified by comparing their performance to human performance in similar tasks. In technology, It is the basis for some of GOFAI, especially heuristic and satisficing algorithms.

   The tasks achieved by these models mostly seem quite contrived and divorced from everyday life (Dreyfus, 1979, 2007). The cognitive models themselves tend to be parsimonious, like a small computer program. An example of how badly these systems fail at being human-like is that we still do not have an artificial controller for a human-like hand with any dexterity.

   I distinguish (see section 3.2.5.3) between these cognitive models that are based on psychological *theory* (in turn based on science, mathematics, computer models etc.) and models based of *subjective descriptions* (point 6 below).

5. **Personal behaviouristic**

   This is the level of recreating external behaviour. A notable example is passive walking robots (Collins & Ruina, 2005).

6. **Personal-subjective** (see Dreyfus's "epistemological assumption" section 2.3)

This level is concerned with the individual humans, *as we see ourselves*, not as a natural phenomenon to be examined externally. Here we are interested in the subjective (see section 3.3.1), without prejudice to age, gender, race, culture or historical time. It would include anything that pertains to *homo sapiens sapiens as such*, without any cultural additions, such as anything which would be specifically western, modern, well-trained or adult. It would include the *facility* to learn any language and cooperate with others, to construct edifices and imagine worlds. This level excludes anything that is culture-specific, like literacy, or any particular system of logic. Our favourite cultural artefacts, language and mathematics, are already amply explored in rational-idealised AI, in the points below. This level is this thesis's *ultimate* goal – anthropic AI. But since we do not have this yet, the current goal has to be making strides in this direction.

7. **Social-behavioural**

   In this level basic cultural artefacts, like language, are explored. This is where generative grammarians (like Chomsky) argue with other schools of linguistics, like statistical linguistics.

8. **Social-normative** (Logic, Bayesian)

   This is the level of *normative* cultural artefacts, like logic, laws, etiquette, etc. Specific to the west are logic, mathematics and science. Much of GOFAI is in this level, e.g. Simon's Logical Theorist (Newell & Simon, 1956).

   One should recall that this level has positive and negative aspects: On the positive side social norms give us science and technology, without which there is no AI (and so many other things). Moreover, without our western normative traditions I could not write this – I would have no Latin alphabet, no computer, no web or email, and no readers. The western scientific tradition also gives us empirical methodologies that we need in any modern systematic pursuit, including technology or AI. On the negative side our tradition gives us the drawbacks of the rationalistic views, as described by Winograd & Flores (see section 2.4), and as critiqued by Dreyfus (see section 2.3). As Minsky had it, AI is stuck – and I suggest it is stuck at this and the biological levels (McHugh &

Minsky, 2003).

Every level seems to require the levels before it in this list, but its existence is a contingent fact about the levels before. Not all functioning organisms have neural systems; not all neural systems we would want to attribute full-blown cognitive abilities to; not every cognitive mind need generate a subjective perspective; it is reasonable to assume that young children have phenomenal experiences even before they acquire any specific culture; and not all cultures developed logic as an articulated body of knowledge, a few developed mathematics, but science as we recognize it today was only developed in the west relatively recently.

These levels can be seen as all relevant simultaneously, as perspectives (see section 1.4.3). However as AI developers, we need to choose at which level to focus our efforts. So far AI technology has mainly been at levels 8 (Logic programming, Bayesian approaches, some of GOFAI), 4 (GOFAI, cognitive simulation) and 3 (Neural nets). Some thinkers, unpopular with the majority (e.g. Dreyfus), would argue that the cognitive level (4) does not exist as such, but is only an artefact of our present scientific fashions (level 8). This deserves some discussion.

3.2.5.3 The cognitive level is problematic

The term “mind” (or “mental”) is used in the literature in two distinct senses. The first is the lay person's intuitive sense, as in “what is on my mind” - it is the subjective world of one's own experience (or consciousness), it is what is accessible by introspection. This is often called the “phenomenal mind”. However, in the psychological (cognitive) literature, the term “mind” refers (often if not always) to imputed processes that go on inside a person in order to achieve the performance that people *empirically* achieve (Nisbett & Wilson, 1977, p. 232). For example, much of cognitive science would say things like “the mind is what the brain does” (Skinner, 1987) or “the mind involves top-down processes” or “the mind has short and long term memories”. The first notion of the mind is subjective, while the second notion *aims* to be “objective”, but is actually (at least for now) speculative, with some correlational backing. We call this second notion the “cognitive mind”, but I beg to not make this usage a claim that these mechanisms exist other than as a level of analysis.

Another way that the cognitive mind is often defined as "*the underlying mechanisms – the cognitive processes and structures – that give rise to … effects*" (Ericsson & Simon, 1981). One can only presume this includes the seemingly universal behaviour of reporting on the existence and content of a phenomenal mind. Miller says (quoted in (Nisbett & Wilson, 1977)) that "*It is the **result** of thinking, not the process of thinking, that appears spontaneously in consciousness*". So also the word "think" seems to also have this duality – here Miller defines thinking as the *underlying* process that we do not experience, while in everyday speech (and introspection) we usually consider "think" to refer to things we can be conscious of, as in "I think that *this* path to the café is shorter".

Note the aim of psychology (as a science) is to make cognitive models that reflect accurately what is going on in the brain, and therefore predict accurately not only externally-observable behaviour but also subjective experience (Seth, 2010). I have no argument with this aim, but I must point out that we are probably at least decades away from any models that in any way predict human behaviour and experience outside of constrained lab situations. So for now, the "cognitive mind" of which psychologists talk is no more than a set of theoretical constructs, with *some* evidence. Therefore, treating the cognitive mind as reflecting any sort of system that we can usefully emulate in usable AI is only one option, and one which has arguably already been exhausted in terms of AI technologies. The main thrust of this thesis is promoting the use of the subjective mind as a source of ideas for AI, via introspection.

3.2.5.4 Simultaneous multiple levels in computers

Unlike the human mind, where there are no levels to be found (see section 3.2.5.1), in a computer there *are* levels, since computer were *designed* that way. And yet, the question of what a computer is doing is not clear-cut. One could say it is working, if the power supply is on. One could say that nearly all its circuits are dedicated to keeping 0's distinct from 1's (B. C. Smith, 2005). One could view it as a 64-bit processing machine, shuffling chunks of data that size. One could view it as running MS-Windows or Linux, or as running some software package, say an "Oracle" database. One could also see it as running some application like a billing system, or as "collecting debts". Less positively, one could see such a machine as perpetuating the injustices in society. While within a specific discourse such as "what operating system is being run" there are specific

answers, that can be categorically true of false[8], I can see no way of determining what is the "correct" description in general, perhaps except asking what was the intention of the person configuring the system in the particular way that it is - at a particular moment. Even this definition may be incorrect, as a learning AI system may bear little resemblance to any intention by a human. My cat will always view a computer as a source of heat to sit on. Again, there is no "correct" way to view a computer. See section 1.4.2.

So even when we say that a computer is running some non-rationalist AI system, at the same time we can see that the computer is keeping the 0s and 1s distinct, is checking various checksums, and is running like a very digital and logical computer. The level in which a computer would be running a non-rational system is at the level of *our intentions for it*. We could call that the "conceptual level" or D.C. Dennett's (1989) "design stance". Every algorithm can be implemented by any universal machine, and a universal machine can be built in many ways, so I may argue against Bayesian statistics or logic, but still run my anthropic algorithm on a system that used Bayesian statistics (perhaps) and boolean logic (for sure) at another level. This topic relates also to pan-computationalism (Müller, 2009). Related points are found in sections 6.3.5.3, 6.3.5.4.

3.2.6 Anthropic AI so far

I see less **effort** being invested in human-like AI in comparison to other technologies, and very little effort in teaching this field (see section 3.2.1), the main effort that *is* ongoing in creating human-like behaviour is chatter-bots for commercial (Deryugina, 2010) and Turing-test (Mladenić & Bradeško, 2012) purposes; this is human-like, but not anthropic. Moreover, I am aware of no *technology-oriented* efforts to create AI that emulates any *subjectively-experienced* mechanisms of the mind (though Agre tried emulating Heidegger's phenomenology, see section 4.3.2). On the borderlands between philosophy and science, machine consciousness is a vibrant effort (Gamez, 2008). Most efforts to create human-like technology apply tried and tested paradigms of the rational-idealised kind, such as machine learning. This mismatch of means and ends invites Dreyfus's (1979, p. 100) quip about trying to get to the moon by climbing a tree.

8 Though note the further layers of complexity introduced by virtual machines or various types, and nested configurations thereof.

An interesting point is that historically "*cognitive anthropology was nipped in the bud in the early 1970s*" (Boden, 2008, p. 32). **Margaret Boden** (b. 1936) has a whole chapter (8) entitled "*the mystery of the missing discipline*" about how anthropology is ignored in cognitive science. This absence of an anthropological angle to AI could speculatively be attributed to some general abhorrence of the primitive, as testified by the observation that Herbert Simon (see sections 2.2) has contributed to "*every social science discipline except anthropology*" (stress added) (Augier & March, 2001).

Anthropic AI assumes a pragmatic distinction between the layer of intelligence that allows for human learning on the one hand, and the content of such learnings (culture) on the other hand. I find only two AI efforts with a technological orientation that were made that are relevant to anthropic AI in that they share this assumption:

The first is **CYC** (Lenat, Prakash, & Shepherd, 1985), which was an attempt to give a computer the rules, or knowledge, underlying "common sense". These rules were done in the spirit of an expert system, and the goal of the project was to use this "common sense" to overcome the brittle nature of such expert systems. However (as (Dreyfus, 1996) typically has it) "*Lenat predicted that in 10 years Cyc would cope with novelty by recognizing analogies and would then be able to teach itself by reading the newspapers. Time is up and he seems to have made no progress on this front*". In a sense there were *three* layers envisaged here: the innate expert system engine (pre-programmed), the rules to be fed in, and the culture to be accumulated after such rules started functioning.

Another attempt to do something comparable to anthropic AI was **COG** (Brooks et al., 1999). The idea was to build on Brooks's older insect-like intelligent system, by attaching a system of little more internal sophistication to "*an upper-torso humanoid robot called Cog*", and to train it over time by interaction with the environment, like human infants develop. Again, Dreyfus puts it in his accurate but cruel style: "*the 'long term project' was short lived. Cog failed to achieve any of its goals and the original robot is already in a museum*" (Dreyfus, 2007).

There is a distinction (key to debugging) between these efforts, that can also be applied to AI in general: COG, like neural nets, cannot explain its behaviour in (reasonable length) language-like communicable form. On the other hand, CYC, being an expert system extension, can print out a trace of how a logical deduction was arrived at. This

explicitness and clarity is important in two senses: humans, at least once they acquire the sophistication of very few years of age, can articulate reasons for (at least some of) their actions, and as will be argued in section 5.2 these are not “noise”. Additionally, having the ability to explain how some outcome was arrived at is invaluable in debugging (see sections 3.3.1.1, 7.6.3).

3.2.7 Knowing that vs knowing how, and a hint on data structure

Let’s preliminarily look at the kinds of data structures involved is different approaches to AI. The main data type of AI systems often reflects a view or a perspective about what type of knowledge is most basic, and should be the native data type of the AI software. A word is due about this and some other arguments in philosophy of mind: when these arguments are being conducted they are most usually about what is the *correct* way to *understand* humans, *scientifically*. However, as technologists, and even more so as technologists looking for ideas, we have no need for correctness, since we are not doing science (see sections 4.2.4, 4.2.5). We need perspectives, *ideas for technology*. So if in some scientific or philosophical argument scholars are adamantly arguing between two or three positions – in AI-as-technology we can try each and every one in turn (more about this laissez-faire attitude in technology in sections 1.3.2, 1.4, 1.5, 4.2.4, 4.2.5). This section does not make claims that will be part of the main argument at this stage, it is an introduction to sections to come (5.2.2, 7.5, 7.6.2).

As Fantl (2014) describes this well-known distinction, we can distinguish three kinds of knowledge:

1. knowing how to do something—say, ride a bicycle. Call this “knowing how”.
2. knowing a person—say, your best friend. Call this “familiarity”.
3. knowing that some fact is true—say, that the Red Sox won the 2004 World Series. Call this “knowing that”.

Following Fantl in ignoring (2), which seems to be mere recognition, and bearing in mind our purpose of producing novel human-like AI, let us examine the possible positions on the relations between “knowing how” and “knowing that”. Note that the psychological distinction between explicit and implicit knowledge is similar though not identical to the knowing that/how distinction.

In terms of AI the three positions have different implications for design. Considering one's stance on the basic knowledge type, we should also consider what would be the AI system's basic data structure.

1. **Intellectualism** is the position that all knowledge-how is based on (and reducible to) knowledge-that. This is the position reflected by Simon's GPS and many other projects, including prolog, expert systems, and arguably the vast majority of classic (and statistical) AI (Dreyfus & Dreyfus, 1986, p. 146). When people design AI systems they often discuss (explicit) knowledge about the state of affairs in the world, rather than skill. This position is most explicit in symbolic-AI (like expert systems) or more subtle in the statistical knowledge of learning systems.

   Appropriately, the data structures found in classic AI are usually combinations of the following:

   - Statements of facts and/or of rules, as in first-order predicate calculus.
   - Statements of probabilities, in Bayesian and other statistical systems.

   An interesting point is that the "naked" expert system as such is like a prolog interpreter, an infrastructure for inferences and a store of explicit "knowledge-that". A "knowledge engineer" translates human knowledge into this format, sometimes even trying to grasp knowledge-how. Fuzzy logic (see section 7.1) was introduced to help with the difficulty of formalizing skills.

2. **Anti-intellectualism** recognises the existence and validity of both categories of knowledge. This implies that both "knowledge-how" and "knowledge-that" are underpinned by some third terminology, but we have scant idea what that third terminology might be (short of simulating an entire brain) and this will surely be very complex. One could present intellectualism (above) as saying that the underlying uniting mechanism is some sort of "knowing that", while radical-anti-intellectualism (below) can be presented as identifying this underlying mechanism as being sort of "knowledge how". This middle position does not specify what the underlying mechanism might be.

   Pragmatically, we have no such underlying concept which is readily

implementable in technology, at a level that even aims at convincing human-like AI. The nearest attempts are vaguely based on brains (probably since brains are the only place we see an underlying mechanism for both types of knowledge).

The data structures associated with this approach are brain-inspired, mainly:

- Neural nets
- Brook's emulation of multiple cooperating mechanisms (Brooks, 1991)

1. **Radical anti-intellectualism** claims that all knowledge-that is based on knowledge-how, for example knowing that 2+2=4 is seen as just a set of behaviours one can be skilled at, like the ability to say "two-plus-two-is-four" and to produce other behaviours that "apply this knowledge" in appropriate times (Fantl, 2014). This position has not been explored in AI yet, and seems compatible with phenomenology and Dreyfus's works (1979, 2007), especially the emphasis on skill as the basis for a more phenomenologically correct AI (Dreyfus & Dreyfus, 1986). As this thesis aims to open up new ways of conceiving of AI, one should note that this is a "road less travelled by" in AI. It may well be worthwhile remaining sympathetic to this position. Further discussion will be found in sections 7.6.4, 8.4.2.
   A data type that would implement knowing-how needs to have less categorical elements than first-order predicate logic, so that skills can be implemented correctly not just in the canonical and clear situations, but also in *similar* situations. Fuzzy logic (see section 7.1) makes a start, and more advanced examples are provided in the rest of chapter 7.

Another reason to be sympathetic to knowledge-how as the basis for AI systems is found in noting (following Ryle as quoted by (Fantl, 2014) and (Carroll, 1895)) that knowledge-that is inert – it is like statements written on paper or stored in a computer. In order to *apply* knowledge-that one needs to know *how*, but moreover one need to have a mechanism that not only *knows* how to implement such inert knowledge-that, but actually *does* it – like a CPU not only "knows" how to run programs, but actually *does it*. So knowledge-that just cannot function alone in the world, it needs an active mechanism with the right know-how.

Since we are here interested in *anthropic* AI, as distinct from the AI that imitates "western, modern, well-trained adults", we can leave "knowledge that" to be developed in the enculturated, learned phase. We needn't focus on "knowledge that" directly at all, leaving us to focus on knowledge-how. Later (section 7.5) we will see an example of an

AI design where repetition of sequences of actions is the basis of skill-development.

3.2.8 Metaphysical non-problems

There are some debates that spring to mind at this point, but that have no technological impact and therefore it is best to remain agnostic about these. These debates often overlap and are sometimes simply the computer-science vs the philosophical names for similar if not identical issues. The main points of these debates can be summarized thus:

Some technologists would object that chatter-bots are not "real AI", that their mechanism is just a "bag of tricks" that would not constitute "real intelligence" or "a mind" under any construal (Deryugina, 2010). As a programmer, it is difficult to read the source code for "Eliza" and come to any other conclusion. I would like to refrain from making a judgement that *all* bags of tricks are somehow categorically "not a mind", just because "Eliza"'s bag is nearly empty. Who are we to assume that our native intelligence is anything *more* than "a (larger) bag of tricks"? In a sense, in Anthropic systems, like in COG, we are trying to build a system with a base intelligence, that will collect more skills as it learns and evolves – so to speak "filling up its bag" as it goes along. Eventually the behaviour could be sophisticated, hopefully even human-like.

Searle (1980) defines "Strong" vs "Weak" AI. His argument against strong AI is based on the following assumption:

> *Intentionality in human beings (and animals) is a product of causal features of the brain.*

He proceeds to use the "Chinese room" argument to show that

> *Instantiating a computer program is never by itself a sufficient condition of intentionality.*

Searle discusses several replies to his argument, one of them being the "other minds" reply. It argues that we have no way of knowing whether a *person* understands Chinese, other than by their behaviour. So if we attribute cognition to people we must attribute it also to machines that display similar behaviour. Searle says:

> *This objection really is only worth a short reply. The problem in this discussion is not about how I know that other people have cognitive states, but rather what it is that I am attributing to them when I attribute cognitive*

states to them...

But for us as technologists the last thing we care about is “what it is that I am attributing”, or what the *true* meaning of cognitive states is. We care about machines that are fit for purpose, that work well enough. This is all that we will be attributing to our AI systems, and the way to verify the criterion of “fit for purpose” is empirically.

The philosophical or scientific question of what “true” intentionality or cognition is may take many decades to resolve. As technologists, we needn’t wait for that. The problem of whether a mind is “real” or not is a question of philosophy, not of technology – and here we are interested in technology.

Once we have far more advanced anthropic AI systems, the question of how “real” an AI mind is could be re-phrased in terms of the depth of similarity between the artificial and the natural-human intelligences. These days are far in the future.

3.2.9 Ethics

If and when we would have human-like AI, it would give rise to a bevy of philosophical problems. As mentioned above, the strong/weak AI argument is analogical to the “other minds” problem, which is unsolved, by most reckonings (Hyslop, 2014). Moreover, a truly human-like AI would give rise to (at least) three types of ethical problems:

1. Should we treat the AI as an entity capable of *real* experience (for example suffering), and therefore an entity towards which we would have moral obligations ?
2. Would such AI be dangerous, in that it could become aggressive or try to take over the world in some dangerous (to humans) way?
3. Would it be fair to have humans relate humanly towards a machine, and develop an emotional bond with such machines (Whitby, 2011)?

Since human-like AI is in such a pitiful state at the moment, I view it as morally safe to exclude any ethical worries from *this* thesis. This however must be revisited if human-like AI becomes significantly more successful than it has been so far.

There is however an opposite angle on human-like AI. The whole idea of making artificial humans is precisely in order to replace human labour. In a sense we would like,

if possible, to bring back slavery, without any genuine human suffering. Slaves understood our language and customs, and provided personalized service. However, the very mention of slavery is near taboo in our society: we no longer keep slaves explicitly, but we should broaden our view. Human-like AI would possibly alleviate a lot of the tedium of current workplaces, some of which (especially in poorer part of the globe) may seem to future generation as repugnant as full-blown slavery seems to us today.

So there are ethical risks, but also possible benefits in human-like AI. Because of the pitiful state of the technology, none of these will be further discussed.

3.2.10 Anthropic AI: summary

When I use the term "anthropic AI" I mean human-like, pre-cultural AI, aimed at technology. Not AI trying to be a western, modern, well-trained or adult, and not directly aimed at understanding humans per se (as in the science of psychology).

Human-like AI tries to simulate human intelligence. Once we remove the normal biases of the western image of what human intelligence consists of, we get anthropic AI.

Anthropic AI is in a sense human-like AI taken seriously. Humans are born immature, and learn general and specific skills. The general skills can be seen as a maturation of the basic abilities, but the culture-specific skills are, as defined, specific to the environment of the individual. If we want human-like intelligence, then we need this ability to adapt. Most existing learning systems are rational/ideal, and the few systems that aim at human-like behaviour (like COG or CYC) are very far from achieving this purpose.

3.3 Terms of this thesis: "introspection"

This section aims to clarify (to the degree necessary) the very concept of introspection, and to prepare the ground for chapter 4, where I argue that introspection can be legitimately used in some scientific and technological contexts, chapter 5, where I argue that introspection is a promising basis for anthropic AI, and chapter 6 where I detail the kind of introspection recommended for anthropic AI.

Introspection is a type of self-observation. Being the observation of one person, it is subjective, but being on observation of one's own mental states/processes, it is inaccessible to others *in principle*, and is therefore doubly subjective. Phenomenology

(in contrast to introspection, see section 3.3.1.5) (Gallagher & Zahavi, 2012, pp. 28–29), is an attempt to study subjectivity by introspection, but in a controlled and peer-reviewed way. In a sense it is an attempt to create an inter-subjective literature based on introspection, an “objective” description of the subjective human condition.

3.3.1 Studying subjectivity

Subjectivity is not a favourite topic of science, for evident reasons: Few of science's methods (empirical repeatability, mathematics, induction, Occam's razor) work in the subjective realm. This area is so fraught for scientists that several attempts have been made (most prominently by the Vienna circle and by Watson (1913)) to banish subjectivity from any discussion, that is, to *legislate* for systematically ignoring what is clearly always there, in each of our experiences (recall Descartes' “Cogito”) (Seth, 2010).

3.3.1.1 Why subjectivity?

As we saw in section 3.2.5.2 there are many levels in which we can discuss and/or simulate humans. Some of these are impractical (e.g. the atomic level) Some have already been tried repeatedly in AI (cognitive simulation, neural nets, mathematics, logic, probability theory). Strangely, the level most available to us as individuals, our own subjective experience of ourselves, has been neglected. Below are some of the reasons to choose this level (beyond just trying something different).

Since our concern is with a technology that aims specifically to behave in an anthropic manner, an intuitive preference could be to use **terminology and mechanisms** that people can readily relate to. Humans relate well to each other's subjective narrations about their internal mental states: many conversations start with “How are you” - soliciting precisely this sort of narrative, and some continue with “How could I do X”, soliciting instructions based on the respondent's internal understanding of herself (for an extended discussion of how humans instruct each other successfully using introspection see section 5.2). Moreover, humans who have no language in common still assume the existence of phenomenal mental states in strangers in terms familiar to their own daily discourse. One of the benefits of the approach of this thesis is that it is plausible that if we were to build systems based on our own subjective daily experience, such systems would have a good chance of being easier to relate to, and hence more useful

(eventually) as, for example, caregivers for the elderly (see section 3.2.2.3).

People like anthropomorphising things, like “the tree wants the sun” (see section 3.2.2). In order to allow people to interact more smoothly with robots, we need to have some level in the robot that functions in a way similar to humans, so that both the robot can parse how people are behaving, and people can parse how the robot is behaving (see section 3.2.3).

A secondary reason to be interested in the subjective is that simulation of subjective processes would be easier to debug than lower level processes, since we have no intuitive understanding of lower levels such as our own neurons or neuron-assemblies. Going any lower than subjectivity would therefore complicate debugging.

Going to an any higher, more formal levels (where cultural assumptions begin) lead us into methodologies that tend to be clunky, possibly infallible, and not truly human-like.

3.3.1.2 Locating subjectivity

Most AI systems are implemented in software, on a computer platform. The computer platform is designed to implement a formal system (though being a physical object in the world any computer is subject to the whims of real world, such as power cuts) (B. C. Smith, 2005). On such a formal system, there is usually an operating system and many other pieces of software all making no attempt to break this formal structure. Actually, it would be in-principle impossible to break out of a formal system: that is why we use *pseudo*-random number generators – there is nothing random in a formal system[9]. However, using software, one can try to simulate non-formal systems, such as the weather, to a certain resolution. Such is the type of AI that this thesis promotes: using software for simulating the processes that we experience subjectively, the human thought process, which is not formal. However, a human can learn to think formally (e.g. in school), and so in principle human-like AI could also learn to use logic and mathematics, thus creating a 3-layer “sandwich” – formal systems below and above, and the mess of actual human thought (Goldie, 2012) in the middle.

Consider these three levels in humans: The lower level, the hardware, which could be

9 A truly random value can be obtained from outside the computer, either by using keyclick timings (as is done in many Linux systems) or inputting some quantum-value from a special device, obtaining a truly random value. In any case, the randomness is being imported from outside the formal system (Isensee, 2001).

called the implementation level, and is well understood in computers, and is being painstakingly researched in humans by neuroscience. This level in humans is not accessible to consciousness (Nisbett & Wilson, 1977), and therefore is not readily available to the AI developer. The middle layer, where we have the informal "mess", is available to our subjectivity on the human side, and is called for in this thesis as software on the computer side. The upper cultured level is achieved by years of education in humans (imperfectly, since humans are fallible (Ariely, 2009)). We could try to train introspection-based AI systems to do formal tasks, but that may be a fool's errand – we can use existing formal systems for that.

The argument against much of AI is that it tries to short-circuit this sandwich, and pretend that humans are rational through and through, this is most pronounced in logicism (Bringsjord, 2008), but is visible also in Simon's work, e.g. the General Problem Solver (Newell & Simon, 1961b). Some strands of cognitive science try to see the "mind as machine" (as per the title of (Boden, 2008)) ignoring the messy middle, at their peril. Other parts of the cognitive science community, such a Papert as quoted in section 1.1, recognise this messy middle, and very few if any have tried to simulate this technologically (see COG as the nearest attempt in section 3.2.6).

3.3.1.3 What is subjectivity

So what is subjectivity? It is the fact of how things look *for us* (Seth, 2010). For us collectively, and/or for each of us as an individual. In the present, and/or in general. The subjective world, a bit like the objective universe, seems **endlessly complex**, but worse, it seems that we can never agree on anything in the subjective realm, so no division of labour is possible, no proper gathering of data, no science, and possibly no systematic study at all. The subjective realm is not made of "*moderate-sized specimens of dry goods*" with which out mind is so adept (Austin & Warnock, 1964, p. 8). These difficulties are why it is often called "ineffable", and left to the poets.

The idea that studying subjectivity is difficult or impossible is the received view in most of the English speaking academic world. Strangely, where science fails many other professions succeed: Lawyers convict or clear criminals with arguments that discuss intent, feelings, etc. journalists discuss the emotional states of politicians, businesspeople and other news-makers, and novelists have little problem with

discussing the subjective. Scientists will protest that their "findings" are non-repeatable, non-quantifiable, consist of "folk psychology" (Ravenscroft, 2010) etc., but that does not stop these professions from being consistently successful on their own terms. Moreover, there have been at least two major attempts to explore subjectivity systematically outside the English speaking world. One was in **Indian** philosophy (The classical introduction being Zimmer, 1951), which is outside the scope of this discussion (but surely should be explored in the context of AI elsewhere, see section 8.4.5), and the other is **phenomenology** (Gallagher & Zahavi, 2012) (see section 3.3.1.5).

Note again that in many academic discussions of subjectivity the accepted nomenclature (which I will follow here) for the environment is either "physical ***universe***" or "human ***world***", to denote the objective and the subjective perspectives respectively.

3.3.1.4 Subjectivity can be studied

An important part of subjectivity is perspectival-ness, (apparently) another is the qualia, or the what-it-is-like-to-be a subject in a situation (Nagel, 1974). I take no stand on the anatomy-of-subjectivity – that is a philosophical discussion with little or no impact on technology (Mandik, 2001). This section shows that subjectivity *can* be explored, in technologically meaningful ways.

Some rudimentary starts on exploring subjectivity in cognitive science have already been made in cognitive science. Subjectivity seems endlessly complex, so one should be careful never to "tick the box" labelled "subjectivity" and consider any one example definitive. The following are all good starts:

- Any perceiving agent, even a camera, has its own ***geometric point-of-view***, and need not take a god-like objective perspective. So (for example) humans need not calculate (objective) motion equations in order to know how to catch a ball, but can implement some preferences regarding the angle of sight *from their own point of view as a player*, and catch a ball with minimal difficulty (McLeod, Reed, & Dienes, 2003).
- Every system that tries to make sense of a situation (rational or not) has a limited amount of information, computational resources, and time available to it - this is

Simon's "**Bounded rationality"** for which he got the Nobel prize (Nobelprize.org, 1978; Simon, 1996a). We are "only human" (see section 2.2.1).

- All learning AI systems can be seen as subjective, in that every running specimen of a machine-learning algorithm is a product of its own training set, in a sense a product of its own life-experience. The field of machine learning is acutely aware of this in how it manages training sets.

As we see, subjectivity is not one thing, but a target for a (possibly never ending) search. It is a search to be pursued (at least) for as long as the search is fruitful.

Another avenue to explore the subjective would be introspection. We all have a direct (Hyslop, 2014) and relatively unhindered access to our subjectivity in introspection. Each and every one of us is a specimen of human subjectivity. Another way to study subjectivity is to read reports given by others (perhaps from the phenomenological tradition), but these are also based ultimately on someone's introspection.

3.3.1.5 Phenomenology, hetero-phenomenology

As we saw in section 2.5.1, "*Phenomenology is the study of structures of consciousness as experienced from the first-person point of view*", and is arguably as old as Buddhism, but (at least in the west) it "*came to full flower in Husserl*" (D. W. Smith, 2013). Heidegger, Husserl's student, revolutionised the ontology implied by phenomenology – for Husserl and most of his students questions of being or ontology were "bracketed" or set aside, leaving the phenomenologist with essentially an idealist ontology (D. W. Smith, 2013). Heidegger (see detailed introduction in sections 2.4.1 and 2.5) argued that we cannot understand the human condition other than in the human's concerned involvement with the world, and this human interaction is Heidegger's new notion of an ontological foundation. In making this human condition the foundation, Heidegger points out that fundamental to a human's interaction with the world is interpretation - making sense of the situation one is always already inside.

Phenomenology is the systematic and peer reviewed study of introspection (Gallagher & Zahavi, 2012, pp. 28–29), but it is also a literary tradition, including Husserl, Heidegger, and others. Some would argue that it is more akin to a sect, where Heidegger's musings are accepted as gospel, than to a truly peer-reviewed and debated

discipline (Romano, 2009). Conveniently, I do not need to have a position on this matter, as I am *not* promoting a phenomenologically-correct methodology for AI (like Dreyfus), but *individual* introspection by AI practitioners. Phenomenology *is* important to this thesis because it is one of the main alternatives in the literature to cognitivism and it is similar to my argument in that it stresses the subjective. In an important sense this whole thesis aims to pave a road about half-way between phenomenology and classic AI - agreeing with the phenomenologists that rationalism is limited, and we need subjectivity, but breaking with phenomenology and moving to the side of classic AI in demanding that software be written, regardless of how much violence that may cause to our models of subjectivity. Elegantly constructed models, as phenomenology has (which are not programmable) are useless for us as technologists (Dreyfus, 2007).

Hetero-phenomenology is a systematic attempt to explore the subjective "in the second person" using interview techniques, hence "hetero" - "other" (D. Dennett, 2003). An interesting sub-case of hetero-phenomenology is Hurlburt's (2011) effort to explore "pristine experience" using scientifically valid practices, and maximum care (Hurlburt, Heavey, & Kelsey, 2013). For example, in some experiments subjects were asked to carry a buzzer, and report on their experience when the buzzer goes off, immediately, so as to minimise pollution of the introspective input from later thinking. The focus of this work is understanding our actual experience, an illusive subject-matter, with as much honesty and rigour as possible.

3.3.2 Defining introspection

Most of the rest of this section (3.3) will examine some definitions and delineations of introspection, and discuss how they stack up.

This thesis deals only with introspection within the context of it being a plausible basis for AI development, and so has little need to involve itself with the many debates about the nature of introspection itself.

Let's start with a definition and a characterisation of introspection:

Overgaard (2008) defines introspection as:

> *an observation and, sometimes, a description of the contents of one's own consciousness*.

I will assume in the rest of my discussion (with the bulk of the literature) that indeed consciousness is what is enumerated in the observation process known as introspection, so when you introspect you are looking at consciousness. This means that I also assume there is no other consciousness (non introspectible) and no other introspection (which is not observing consciousness).

Schwitzgebel (2012) surveyed many definitions of introspection and gives 6 characterizations of introspection. Most definitions of introspection include the following criteria (abbreviated):

1. About **mental** events, states, processes, etc.,
2. About the **first-person**,
3. Simultaneous or in **temporal proximity** to the mental event, state, or process (not a medium or long-term memory).

By most definitions introspection also:

4. Is **direct**, not involving (at least any complicated) inferences,
5. Is detecting **pre-existing** mental events, precesses (etc.),
6. Requires an **effort**, not constant or automatic.

For building AI, the targets of introspection (at least initially) would be *processes that influence any operations on information*, that can feasibly be replicated in a computer – as opposed to the vague ebb and flow of subtler emotions, levels of alertness, and other observable mental states and processes. These later elements may be of use in AI in the future, but that future is further away and does not concern us here. Introspecting the fact that one feels cold, or believes some peculiar fact would have no immediate utility for AI design. The “products of introspection” would be some reports on how information is processed, that would be useful for AI development (see section 6.2). An example of introspection is given in section 3.3.4.

A word is due about the “effort” (point 6 above) required (by most philosophers) in introspection, or even the idea of introspection being an action, represented by a verb. In one sense, introspection is one of the most passive actions possible, since the world of

our own consciousness is available to us without much effort, just for the noticing. So "noticing" is required, as per point 6 above. But in a sense we need to make a further effort when introspecting, to be authentic, to tell things as they are and not as we may expect them to be (by our own criteria or by society's). See section 6.3.3. Also if the introspection is being expressed using words, there is the effort in speaking coherently, e.g. not mixing languages.

3.3.3 A boundary between introspection and science collapses

Strangely, even though introspection is presented by mainstream cognitive science as utterly wrong (Ericsson & Simon, 1993; Nisbett & Wilson, 1977; J. B. Watson, 1913, 1920), there are some similarities between "thinking aloud" (TA), promoted by both J.B. Watson and Simon and introspection (Ericsson & Simon, 1993; J. B. Watson, 1920). This discussion is important (1) since TA is a central technique in psychology, (2) since it presents an interesting boundary case for introspection, and (3) since one of the main characters here, Simon, is also a central pillar of the AI community (see section 2.2). Note that Simon nowhere disagrees with Watson's TA technique, he sees himself as only further elaborating it.

The difference between TA and introspection seems to be that in (acceptable) TA the person reporting his thoughts is naïve, not a psychologist, and the content of the report is about some subject matter *other than* psychological mechanisms.

Further below I will show that mainstream psychology's aversion to introspection, and neo-introspectionists' concern for *correct* introspection both share a scientifically-motivated aversion to unexamined inferences mixed into the data of any systematic study. However, further examination of this preference shows that it is naïve and unwarranted: all observation is interpretative, and "clean" data does not exist.

3.3.3.1 "Thinking aloud" (TA) can be seen as introspective

This section is not trying to *establish* that TA *is* introspective as a matter of fact, but just to show that a case can be made that it is introspective, or that it is *arguable* that TA is introspective. Later I will show how TA was distinguished from introspection.

This section is about TA as a technique. This technique was established by J.B. Watson (1920), and was discussed and expanded on most famously by Ericsson & Simon

(1993). My concern is with the currently acceptable practice more than with Watson as a historical character.

Watson was the leader of the behaviourist revolution and it was he who largely abolished introspection as a legitimate technique in psychology (J. B. Watson, 1913) (see sections 2.2, 4.2.1). Considering that Watson was vehement in denying introspection any legitimate role, it would be very odd to find something similar to introspection mentioned positively in his writings, especially from the same years he was running his anti-introspectionist campaign (Costall, 2006).

However, Watson (1920) introduces an idea of "thinking aloud" (TA), *in contrast to introspection*, without clear definitions or references[10]. He states that for starting an experiment using TA "*usually a request is sufficient*" (J. B. Watson, 1920, p. 89). He adds that the subject has to enter into this experiment in "*the proper spirit*" without detailing what that may mean (*Ibid.* p. 92), and states that a "*scientific man is quite willing to enter into the experiment with zest*", again without leaving us any hint as to what is meant by this "scientific" subspecies of mankind (*Ibid.* p 91). He does, however, give a few examples of such thinking aloud.

The longest example Watson quotes is of a colleague who came to stay in an apartment in which Watson "*had rooms*". He challenged the guest to figure out the use of some contraption belonging to the landlords, while thinking aloud. Here is the full protocol as recorded by Watson (round brackets are Watson's notes in the original text):

> *"The thing looks a little like an invalid's table, but it is not heavy, the pan is curved, it has side pieces and is attached with a ball and socket joint. It would never hold a tray full of dishes (cul de sac). The thing (return to starting point) looks like some of the failures of an inventor. I wonder if the landlord is an inventor. No, you told me he was a porter in one of the big banks down town. The fellow is as big as a house and looks more like a prize-fighter than a mechanician; those paws of his would never do the work demanded of an inventor" (blank wall again). This was as far as we got on the first day. On the second morning we got no nearer the solution. On the second night we talked over the way the porter and his wife lived, and the subject wondered how a man earning not more than $150 per month could live as our landlord did. I told him that the wife was a hairdresser and earned about eight dollars per day herself. Then I asked him if he did not see the sign 'Hair-Dresser' on the door as we entered. The next morning after coming from his bath he said, "I saw that infernal thing*

10 This style of writing was normal at the time.

> *again" (original starting point). "It must be something to use in washing or weighing the baby-but they have no baby (cul de sac again). The thing is curved at one end so that it would just fit a person's neck. Ah ! I have it! The curve does fit the neck. The woman you say is a hair-dresser and the pan goes against the neck and the hair is spread out over it." This was the correct conclusion. Upon reaching it there was a smile, a sigh and an immediate turn to something else (the equivalent of obtaining food after search)* (J. B. Watson, 1920, p. 92).

Note that this "thinking aloud" is *not* a case of the "extended mind" (Clark & Chalmers, 1998) – not a case of thinking using external props (like doing arithmetic, saying out loud or jotting down "*carry 1*"). Rather, it is a case of "letting one's thoughts escape through one's mouth" in the everyday process of trying to figure something out.

Alarmingly (for our attempt to understand introspection) this technique could be seen as being close to a case of introspection. Note however that the definitions we are now using are all recent, and Watson wrote 100 years ago, so any problem with current definitions does not reflect badly on Watson historically, but these problems are still problems *for us*. The argument that TA is a type of introspection would stress that *mental contents* are being verbally reported in TA. Recall the definitions of introspection quoted above (top of section 3.3.2).

I believe there is a prima-facie case that TA *is* some variation of introspection, according to the definitions. But it is presented by Watson as an *alternative* to introspection – he says: "*...a good deal more can be learned ... by making subjects think aloud ... than by trusting to the unscientific method of introspection*" (J. B. Watson, 1920, p. 91). Note that Simon, who continues Watson's work on TA (Ericsson & Simon, 1993), agrees that there is a possible confusion here: "T*he use of thinking-aloud protocols as data was sometimes misunderstood as an attempt to revive introspection*" (Simon, 1996a, pp. 231–232), so this is no idle worry.

3.3.3.2 Two distinctions between TA and introspection

Watson's legacy in terms of the boundaries of introspection is problematic, as we have seen above. Further elaboration of TA was one of Herbert Simon's research agendas. Simon says that: "*...no clear guidelines are provided* [by Watson] *to distinguish illegitimate 'introspection' from many forms of verbal output that are routinely treated as data...*" (Ericsson & Simon, 1993, p. 3).

Let's try to examine what the two poles of Watson's (& Simon's) contrasting of TA with introspection are: What they prohibit is the psychologist introspecting in his armchair, coming up with pronouncements about how his own mind works, and generalizing them as a general scientific fact. The psychologist is assumed to be already invested in some course of research in psychology, and therefore bound to be biased. On the other hand, the naïve person, whether a "subject" in an experiment or a scientist who is *not* a psychologist, is assumed to be neutral. If such a neutral person gives neutral reports, just observations, then his "verbal behaviour" is legitimate unbiased data.

Let me propose two distinctions:

1. The *contents* of TA are the *contents* of the thinking process – in the paradigmatic example of figuring out the contraption (J. B. Watson, 1920, p. 92) the text reads "*The thing looks a little like an invalid's table, but it is not heavy...*". In "classical introspection" the contents of the report can be about the *mechanisms* that are used to do the thinking, rather than the content of the thoughts.
2. In the forbidden "classical introspection" it is the *psychologist*, a trained individual with a research agenda, who is making the report, while in TA it is a *naïve* participant in an experimental setting.

It seems that these distinctions are supported by Watson and Simon:

For the **first distinction**, about the contents of the verbalisations, Ericsson & Simon (1993, p. 58) analyse (J. B. Watson, 1920) as follows:

> *It should be noted that the kind of questioning illustrated by* [Watson 1920's story about the golfer, pp. 100-101] *does not refer to to the subject's memory of a specific instance, but to how he thinks he performs activities* ***in general*** *when he is asked about them. Watson made a clear distinction between analytic classical introspection, verbal questioning of a subject, and thinking aloud. His views on the veridicality of the later kind of verbal report were quite different from his views on the first two*" (emphasis added).

So Simon objects to the generalisations that the golfer makes, to the fact that the golfer tries to give non-naïve, analytical comments about his conduct in general. TA would allow him only to verbalise about the task at hand, in the present moment, with no elaborations or speculations.

Later Ericsson & Simon (1993, p. 247) mention approvingly previous research by

Ohlsson who

> *… coded TA protocols to distinguish between heeded thoughts, on the one hand, and introspections, retrospective reports, and communications to the experimenter, on the other hand…. Reports were classified as introspections if the grammatical subject was the speaker (e.g. 'I,', 'my head'); if the verb was epistemic (e.g. 'remember,' 'feel,' 'know') and if the verbalization did not contain specific information about the current problem*.

So any comments or speculation on the "meta" level, and discussion of how-the-mind-does-it, are banned.

Another insight into Simon's position is provided by his quoting approvingly from Duncker, saying

> *While the introspector makes himself as thinking the object of his attention, the subject who is thinking aloud remains immediately directed to the problem, so to speak allowing his activity to become verbal. When someone, while thinking, says to himself 'One ought to see if it isn't-,' or, 'It would be nice if one could show that-,' one would hardly call this introspection*" (Ericsson & Simon, 1993, p. 60)

So the distinguishing criterion seems to be being "immediately directed to the problem" - again a distinction of content.

Turning to the **second distinction**: is the person doing the verbalisation a professional with an agenda or a neutral, naïve person? Here most of the evidence comes straight from Watson. He says:

> *...a good deal more can be learned about the psychology of thinking by making subjects think aloud about definite problems, than by trusting to the unscientific method of introspection* (J. B. Watson, 1920, p. 91)

His phrase "*method of introspection*" is clearly a reference to "*the introspectionists*" - the psychologists that have not yet been converted to his behaviourism. Later he says that

> *The behaviourist… is engaged in studying the process of observing as it appears in others, where the activity is not complicated by the demands of introspection. … the behaviourist is a natural scientist and makes his observations upon his fellow man rather than upon himself*" (J. B. Watson, 1920, p. 94).

These are actually two arguments - the role of data-source needs to be separated for two

reasons: the need to notice the data (the role of the scientist) distracts from the task at hand, and separation of the agenda-laden scientists from the experimental subject is just good scientific practice, minimising theoretical bias (see also section 6.3.3).

3.3.3.3 Inferences and confusion

But if we set the word "introspection" aside for a moment, we can see a commonality between Schwitzgebel's (2012) criterion No. 4, that introspection be direct, and TA's requirements as discussed above, that the TA report be by a naïve person with no agenda, and be about the contents of the problem at hand, not about the mechanisms of thought. The commonality is that they both do not want any inferences to be already embedded in the data.

Lets look at this once more: For Schwitzgebel, a neo-introspectionist, two characteristics of introspection are that it is immediate (3) and direct (4). If it is not direct then it becomes speculation, philosophy, psychology or one of many other things, but in ceases to be pure introspection. Introspection is Schwitzgebel's term for "good" non-inferential reports.

For Watson, Simon, and other psychologists TA is the thing *without* inferences, the "raw data". They call that which contains inferences, speculation, philosophy or psychological theory "introspection", and that is a ***bad*** thing.

So it would seem that all we have here is a confusion in terminology. They both want the non-inferential, "pure" reports. It seems that we have the same word, "introspection", being used for the same thing, except for the inference part where the same word is being used with *opposite* meanings.

A reasonable step now would be to call Watson's Introspection Intro-W, and Schwitzgebel's Intro-S, sort out the differences and make peace. However, such a peace would be predicated on joining the seeming consensus against inference in introspection. The question we must ask is "Is such a position tenable?" regrettably the answer is no.

3.3.3.4 Non-inferential observation is impossible

The desire to separate observation from interpretation has been ruled impossible both in the analytic and the continental traditions. On the analytic side Bogen (2014) quotes

Norwood Hanson, Paul Feyerabend, and mainly Thomas Kuhn in showing that in normal scientific observation:

1. Which aspects of a scene are seen as **salient** and worth recording varies by the set of assumptions the observer is already committed to.
2. Observers conceptualize what they see in terms of their favoured **conceptual framework**. An everyday example would be how different people and different cultures carve up the palette of colours.
3. The very perception can be influenced by “top down” considerations that are not in the actual world. This is demonstrated by Bruner and Postman's research using playing cards with *black* hearts.

All these worries come from a discussion of external observations in the natural sciences. They would be double and triple as worrying in the more complex case of introspection, where the process is entirely subjective, and the very process of observation may impact the content of any observation much more than in the case of observing an external object.

There is another problem with trying to separate observation from interpretation (and eliminate inferences) similar but distinct from point 2 above: When a person who is conversant in more than one language introspects (or thinks aloud) their mental contents tend to appear in more than one language. Any attempt to communicate intelligibly with another person requires translating and organizing one's thoughts into one coherent language. This process of regularising expression is inferential. Arguably, even monoglots have to regularize their language in a similar way.

On the continental side, one of hermeneutics' *main* points is that all observation includes some interpretation. One quick way to make that point from the continental side is to recall the hermeneutic circle: we make sense of the whole only because we have a sense of the parts, but we also make sense of the parts only as since we already have a sense of the whole (see section 2.5).

Like Schwitzgebel’s (2012) introspection and Watson's/Simon's “thinking aloud”, hetero-phenomenology (section 3.3.1.5) is also an attempt to get at the “raw data” of subjectivity.

3.3.3.5 A boundary between introspection and science collapses: conclusion

Though they come from different apparent traditions, both cognitivism and the new introspectionists view themselves as being within the scientific tradition, and want the "raw data". They want to have "data" so they can then work on the "mechanisms" in an open, peer-reviewable way. The idea is to reach some scientific *truth.* However, for AI we don't need (or want to wait for) such objective, singular truths about intelligence. The scientific/cognitive understanding of intelligence is a long way in the future, and the continental candidate we have for such an "objective" or well-received "truth" is Heidegger (see section 3.3.1.5), and we saw that Heidegger is not readily programmable (Dreyfus, 2007). AI-as-technology cannot wait for these debates to be resolved, nor is such a wait necessary.

As we saw above (section 3.3.3.3) the attempt to separate the "good" from the "bad" observations of mental states has hit an impasse. Without prejudice to the efforts being made in this direction (Gallagher & Zahavi, 2012; Jack & Roepstorff, 2003, 2004), for the current purposes (technological AI) this effort must be abandoned – at the price of stating the following at its starkest: **This thesis will be promoting introspection, of the *bad* type**: Whatever it was that was forbidden, in its broadest form - I am promoting *precisely* that, i.e. self-reflective introspection, about mechanism, by theory-laden individuals, with inferences. Consider even that *some* introspectors may be self-conscious and evil charlatans – this possibility has to be carried forward into the rest of the argument. However, I will show in chapter 4 that for our purposes here even "bad" introspection is a legitimate source of ideas, and in chapter 5 that at least much of introspection is plausible as a basis for anthropic AI. I will return to making distinctions between better and worse forms of introspection in chapter 6.

3.3.4 What kind of introspection is recommended

This section is to a large degree a summary of chapter 6, where the arguments are given in detail. An example text follows.

As we have seen, this thesis does *not* insist on the introspection being particularly refined, correct, or exact, since it is not at all clear if that is possible, and in the context of technology, it is unnecessary.

Leaving aside any commitment to "correct" or "uninterpreted" introspection, unlike various thinkers about introspection in philosophy and psychology, in AI we need to have a commitment to the sort of concrete and clear details that are needed in computer programming. We must try to describe the mental states and processes pragmatically, in terms that may be programmable. These descriptions will most certainly be partial, and we should never pretend that the description is exhaustive, or near-exhaustive. We may also relinquish any search for a quick and efficient way of distinguishing between "good" and "bad" introspective reports, not so much because such gradations are not possible, but because over-attachment to correctness and precision has rendered most previous attempts at phenomenological AI sterile in terms of producing actual testable systems (Dreyfus, 2007). The sort of introspection we should aim for is **mid-depth**: roughly half way between Simon (with his positive commitment to programmability) and the phenomenologists with their commitment to observing subjectivity as it appears, rather than as it should be. Note also that in sections 4.5, 5.3 I will speculate that all programming is introspective. In that specific case the introspection is restricted by the limitations of the programming environment (e.g. python, i386 machine) to a small structured palette. In introspecting for AI, we look at the mind as is operates freely, and only later formalize.

A "piece of introspection" *useful for AI* would describe some sort of interaction with the environment that would ultimately be programmable. Any introspective observations that have no impact on external interaction or are not programmable would be "epi-phenomenal" and have no *technological* significance.

Consider the following example introspective report:

> *How do I do long division? Damn – it's been a while – it was that tall teacher that taught that, right? OK, let's see – you take the number to be divided and put it here near the top of the page, and then there was that angle thing you draw….. I used to like that angle!* [...non verbal recollection of the pleasant "liking"...] *Now where do we put the other number – di-vi-sor, was it, or di-vi-dor? Here? That doesn't look right.…… what was that teacher's name? I really need to get this done before Jim comes in…..*

It shows how irrelevant thoughts (and non-verbal reminiscence) such as "I liked..." intrude. It shows how fears and reminiscence drift in and out of consciousness, and how shaky one's real grasp on issues often is, behind any pretence to be logical (see section

3.3.1.2). In all these respects this example is more realistic than (say) Simon's claim that he "thinks in mathematics". I here assume that Simon also has human concerns, as does the example above. But note also that this example is not as clean and refined as Heidegger's descriptions of interacting with a nail using a hammer. The above shows us something about how we *actually* do things, as far as *we* are concerned, and as Papert's said in the interview quoted in section 1.1 (McCorduck, 2004, p. 339).

Humans can view the same situation from multiple perspectives (see section 1.4.2). Can we program these perspectives? Can we program their multiplicity? If one were a scientist sworn to tell the truth, the answer should be "no", since humans' subjectivity is too complex. But as technologists, this is not a question but a challenge to be answered. Surely the challenge of multiple perspectives, and generally the challenge of **pragmatic approximations** of subjective perspectives is easier than (Dreyfus, 2007)'s pedantic challenge of programming something as nebulous as the sum-total of subjectivity. We can do this step by step. The Phenomenologists may protest that it is "not Heideggerian enough" (Dreyfus, 2007), but Rome was not built in a day, and the worst enemy of the good is a nebulous idea of the perfect. How precisely such research can be done will be the subject of chapter 6.

4 Introspection may legitimately be used for AI

Table of Contents

4 Introspection may legitimately be used for AI....94
4.1 Introspection as “impossible”....96
4.2 Introspection as “forbidden”....97
4.2.1 Watson....97
4.2.2 Cognitive psychology's attitude to introspection....98
4.2.3 Other objections....101
4.2.4 Context of discovery / justification....101
4.2.5 Truth in science vs in technology....102
4.2.6 Example & summary of “introspection is forbidden”....105
4.3 Introspection as “commonplace”....105
4.3.1 Sweeping testimony....106
4.3.2 Specific apparent cases....107
4.3.3 Mainstream cognitive science uses introspection....110
4.3.4 Introspection is “commonplace”: summary....112
4.4 Introspection as “desirable”....112
4.4.1 Introspection & phenomenology....113
4.4.2 The Neisser-Dreyfus debate....113
4.4.3 Introspection vs. phenomenology....114
4.5 Introspection as “unavoidable”....114
4.6 A hybrid position....115
4.7 Types of truth in introspection....117
4.8 Introspection may legitimately be used for AI: summary....121

This thesis argues for introspection as a basis for developing anthropic AI. Having defined some some of the terminology in this claim above (chapter 3), this chapter will argue that introspection is a legitimate source of designs in AI. Those who believe this is a non-scientific approach have misunderstood the different relations of science and technology to truth, those who believe is has already been done are granting themselves full credit for half a step, and those who take great steps into introspection, produce no concrete AI systems.

The possible attitudes to the status of Introspection-for-AI that will be surveyed are that it is:

1. Impossible: Comte and other thinkers considered introspection impossible.

These (by now uncommon) positions will be briefly discussed (and dispensed with) in section 4.1.

2. Forbidden: Herbert Simon and the mainstream of cognitive science object to Introspection quite dogmatically, based on Watson's (1913) paper. I will show that their objections do not hold under the modern analysis of the context of discovery vs justification, especially not in a technological context (section 4.2).
3. Commonplace: Solomonoff (1968) and others admit that much of AI *was* done introspectively, and therefore may consider my main point to be trivial. I will show that so far AI researchers only used introspection in a *shallow* manner, timidly, as if it were illegitimate (section 4.3).
4. Desirable: Dreyfus (1979, 2007) and others are all for introspection (at least within the context of phenomenology) but they do not generally do much programming. This is the main “dissident” group from AI (section 4.4).
5. Unavoidable: A preliminary case can be made that inventing AI without introspection would be impossible, if only because programming requires introspection under a specific role (section 4.5). This (novel) discussion is delayed to section 5.3.
6. A hybrid position: Introspection may *already* be seen as “commonplace” in the context of discovery and “forbidden” in the context of justification. A case can be made that I have said nothing new so far. I examine what the consequences of this supposedly established practise would be if this description were true (section 4.6).

The next section (4.7), surveys various positions regarding the types of truth available in-principle through introspection.

The purpose of this chapter is only to show that my position (encouraging introspection as a basis for designing AI systems) is novel, and legitimate. The next chapter will argue why one positively should *expect* introspection to be a *good* basis for anthropic AI. Following chapters will show some details and examples and will discuss some consequences.

In traversing and analysing the various positions scholars have had towards

introspection-for-AI, this chapter also provides a more fine-grained and clearly-categorized view of the state-of-the-art than was possible in chapter 2. Though the main thrust of the argument is against those who oppose introspection, this chapter will also contrast this thesis' unique position within the field by critiquing other scholars, not least sympathetic ones such as Dreyfus, Agre, and others.

Throughout this chapter it is important to bear in mind that I am not yet arguing for the *benefit* or profitability of using introspection for AI. The main thrust of this chapter is silencing those who would shout down any mention of introspection, those who argue that using introspection is illegitimate, unscientific, or in any other way disallowed. In passing this chapter will also demonstrate that this rehabilitation is necessary, since no AI researcher uses introspection as such without compunction.

4.1 Introspection as "impossible"

As Overgaard has it

> *Brentano argued that a paradox exists* [in introspection] *in the relation between observations of 'inner' mental states and 'outer' objects. In order to observe and know about, say, an experience of a red apple, one must turn one's attention from that outer object which was cause to the sensation. This should logically make the relevant experience cease to exist, thus also the attempted introspection. … Comte's first objection was that one cannot have an identity between the observer and the object of observation in science. He argued that the observer cannot be 'split in two' so that one part observes the other, and, thus, observation of one's own inner experiences is an impossible project*" (Overgaard, 2006).

Later thinkers such as Wundt resolve this problem by saying that introspection is based on a "change of focus" from the outer to the inner, and that in introspective vision is to a large degree a memory, a recollection of the mental experience rather than a direct report.

These objections are of ongoing philosophical interest (Schwitzgebel, 2012), but can be sidestepped for this thesis: We can assume with Wundt that introspection is only a memory; Here we are interested only in inputs for technology design and not in some absolute or even scientific truths (see sections 1.3.2, 1.4.2); The content of introspections exist as a matter of fact (Seth, 2010) regardless of any qualms we can have regrading their temporal status, or admissibility in either philosophy, science or

technology. The notion that we are dealing with particularly precise information in introspecting for technology has already been put to rest in section 3.3.3.4. The idea that introspection is inadmissible the the topic of the next section.

4.2 Introspection as "forbidden"

Practical objections to introspection also have a pre-20th century pedigree. For example, Comte (again) argued that introspection "will generate unreliable and conflicting data" (Overgaard, 2006, p. 630). However in terms of impact on current thinking, the most widely quoted prohibition on the use of introspection is (J. B. Watson, 1913).

4.2.1 Watson

John B Watson (b. 1878, d. 1958) is the most oft-quoted scholar for objecting to introspection, and was the most vehement in his objections (Costall, 2006). He does not mince words in criticising his opponents, to whom he refers collectively as "the introspectionists": "*To make the data obtained by the language method virtually the whole of behavior … is putting the cart before the horse with a vengeance*" (J. B. Watson, 1913, p. 172n), "*It is hopeless for me to get his introspective report*" (J. B. Watson, 1913, p. 172). Watson was forceful in his revolutionary talk, threatening psychology with a schism if his world-view were not accepted: "*… either psychology must change its viewpoint so as to take in facts of behavior, whether or not they have bearings upon the problems of 'consciousness'; or else behavior must stand alone as a wholly separate and independent science*" (J. B. Watson, 1913, p. 159).

Watson saw himself as pushing for better **scientific practise** in psychology (viz. "control experiments" (J. B. Watson, 1913, p. 171)). He contrasts "the behaviourist" (an epithet he uses for himself) with the (old-style) psychologist:

> *... questions arise which I may phrase in two ways: I may choose the psychological way and say 'does the animal see these two lights as I do, i.e., as two distinct colors, or does he see them as two grays differing in brightness, as does the totally color blind?' Phrased by the behaviorist, it would read as follows: 'Is my animal responding upon the basis of the difference in intensity between the two stimuli, or upon the difference in wave-lengths?' He nowhere thinks of the animal's response in terms of his own experiences of colors and grays*" (J. B. Watson, 1913, pp. 170–1).

Note how "color" is replaced by "wavelength" - not only more scientific, but also closer

to physics, the most prestigious of all sciences.

To maintain the unity of psychology and the coherence of the scientific programme (in a way that is compatible with the overall scientific program as understood at the time): "*behaviourism… was an attempt to do one thing – to apply to the experimental study of man the same kind of procedure and the same language* … [as] … *in the study of animals lower than man*" (J. B. Watson, 1931, p. ix). Also "*the behaviourist attempts to get a unitary scheme of animal response. He recognises no dividing line between man and brute*" (J. B. Watson, 1914, p. 1). Since animals have no consciousness that can be readily accessed (even by the introspectionists' lights), we should not even attempt the same with humans.

Within his programme of improving psychology's scientific credentials, his most direct attacks are on introspection and say that the content of introspection is "**obscure**" (J. B. Watson, 1931, p. x), the technique of introspection is unclear and imposes self-contradictory demands (J. B. Watson, 1913, p. 163), its terminology is incoherent even in simple distinctions of sensations (J. B. Watson, 1913, p. 164), (and switching to ad-hominem attacks) its practitioners are effete (Costall, 2006, p. 646) and "insufferably prolix" (J. B. Watson, 1920, p. 97).

Like many thinkers, Watson is often remembered by simplistic slogans, such as "introspection is unscientific". His actual position was both more subtle and more strident than that – but that makes little difference to eventual influence on AI (amongst other disciplines). What has influence is his somewhat-flattened memory, more than the living, breathing, complex person he was (Costall, 2006).

I will respond to these and other objections to AI based on introspection in sections 4.2.4, 4.2.5 below.

4.2.2 Cognitive psychology's attitude to introspection

"Telling more than we can know: Verbal reports on mental processes" (Nisbett & Wilson, 1977) is one of the most cited papers "*in the recent history of consciousness studies*" (Johansson, Hall, Sikström, Tärning, & Lind, 2006). Nisbett & Wilson complain that there is "*... little or no direct introspective access to … cognitive processes*" (Nisbett & Wilson, 1977, p. 231). They define cognitive processes as being

"*the processes mediating the effects of a stimulus on a response*" (*Ibid.*). They contrast the cognitive, "real"[11] level with the contents of introspection, which "*is the **result** of thinking, not the process of thinking, that appears spontaneously in consciousness*" (*Ibid.* p 232, quoting Miller, stress in the original). The amount of evidence that the authors marshal is formidable, contributing to the canonical status of this seminal paper.

-

As long as one agrees with their assumptions, this paper stands very well. However, the moment one tries to examine some of the underlying assumptions it becomes less stable. Note that the authors are writing as psychologists, for psychologists, and see themselves as scientists, like when they conclude that (as in the following quote) this paper truly "buries" introspection as an element of psychological discourse (*Ibid.* p. 233):

> *The accuracy of subjective reports is so poor as to suggest that any introspective access that may exist is not sufficient to produce generally correct or reliable reports.*

From a scientific perspective, that is enough to damn introspection as a source of truths. But does that damn it as a source of models for technology? Note that they demand "*generally correct*" reports. Is that the correct level of truth to demand for developing AI? This will be discussed in section 4.2.4.

They complain, (*Ibid*.) that subjective reports are often caused by a-priori theories, rather than by some genuine observation. True to scientific purity, they abstain from any discussion of whether the data of subjective reports may be a mixture of genuine observation *plus* culturally-accepted theories. Their austere conclusion is that these reports should be ignored.

One could worry, together with Dreyfus (1979) (see sections 2.1, 2.3) that the cognitive level is invented (see section 2.1, 3.2.5.3), and therefore Nisbett & Wilson's (1977) definition of the cognitive level as the "real" level would be alarming. But we can let that issue rest, and accept multiple perspectives: perhaps the "real" level is cognitive, perhaps neural, perhaps phenomenal. My only comment on this is that those who hold these different perspectives all seem to be united in believing that there is only *one real truth*, and that all others are utterly mistaken. This may be a reasonable stance in

11 "Real" by Nisbett & Wilson's (1977) lights, but recall the usage of "real", section 1.4.3.

science, but I argue (section 1.3.2) that it is counterproductive in technology. Later (sections 4.2.4, 4.2.5) I will present arguments that deal conclusively with any worries to do with the validity of introspection for our technological purposes.

There is a somewhat entertaining point that will become relevant in section 4.3.3: Having given ample evidence that introspection is of little value in terms of the "actual" "cognitive" processes,The authors move on to speculate on the causes of people's (misguided) confidence in their introspective interpretations of their own thought processes. The authors discuss the "*conditions that give rise to introspective certainty … Confidence should be high when the causal candidates are (a) few in number, (b) ... (e) … In fact we appeal to introspection to support this view.*" (Nisbett & Wilson, 1977, p. 255). Fascinatingly, they appeal to *introspection* as their source of evidence for this list, so perhaps even they are not as averse to introspection as they claim to be.

They introduce a distinction between "intermediate output" that *is* available to introspection, and the actual process, which is not. As an example of how the intermediate results are the contents of the introspection, they quote a person describing how he recalled his mother's maiden name by recalling his uncle's surname (Nisbett & Wilson, 1977, pp. 255–6). This, again, will be relevant in section 6.3.5.1. Let's preview the main point from there, as it shows how the we gain crucial "cash value" from the distinction between science and technology (section 1.3.2): If we know only the outline of *what* is accomplished without the *how* (as the authors complain), in AI programming we can substitute whatever technical trick we have (in our skills as programmers) to achieve something similar in a computer. AI is not science (but technology), the trick we use need not be the same trick that the brain (or cognitive system) uses.

The authors speculate on various reasons why people believe in their own introspections. The last reason they suggest, to sum up their paper, is that it may just be insufferably "frightening" to think that we know nothing of our own mind. I would retort that following Watson (mainly 1913), it had become fashionable in psychology to be harsh on any subjectivity (Costall, 2006). Only recently is scientific psychology recovering from this bias (Seth, 2010).

See also my analysis of "thinking aloud" and how it relates to introspection, section

3.3.3.

4.2.3 Other objections

Herbert Simon (in his AI research, and also in papers on other topics, not least psychology) continues Watson's objections to introspection, though his case is a bit more complex (see sections 2.2, 4.3.3).

If my discussions with various members of the university of Sussex are any indication, the main reason that people in cognitive science object to introspection is various versions of the worry that introspection is not objective, it is impossible to know who is right in an argument, and it is not even internally consistent in that the same person can come up with contradicting reports. I now turn to deal with these and the above (sections 4.2.1, 4.2.2) objections.

As a reminder, this chapter aims to convince the reader that using introspection as a basis for developing AI is legitimate, i.e. it needn't be forbidden. A subsequent chapter (5) will argue that introspection should positively be pursued for anthropic AI.

4.2.4 Context of discovery / justification

There is a distinction (in philosophy of science) between the “context of discovery” where scientists get their ideas, and the “context of justification” where scientists should provide evidence to support their claims (Schickore, 2014). In the context of discovery, Newton was permitted to take inspiration from falling apples when discovering the theory of gravity, and Kekulé was right to dream (while dozing) about snakes eating their own tail, giving him the idea of circular molecules (Rothenberg, 1995). The scientists have to produce their evidence for their equations or models later in the context of justification, but in the inspiration of their discovery they are *entirely free*. Why then should AI researchers not be free, when inventing new AI designs, to use introspection? Before we claim that the new designs are good, we would have to test them as software, so we can be assured that no harm is done to empirical integrity. I am not arguing for letting introspection into the holy-of-holies of scientific fact (Watson (1913) would rightly shudder), only into foyer of scientific ideas. Moreover for the purposes of AI we need even less, just let introspection provide *technological* ideas.

So in principle, all inspirations and sources of ideas are allowed, in the context of

discovery. But there is a seeming problem here, and a subtlety to be noticed. The seeming problem is that this argument trivialises introspection and supposedly says that recommending introspection is as good as recommending taking a long walk or drinking some juice to improve creativity in AI. The crude response would be "so be it", in this section I am only combating the notion that introspection is *forbidden*, and if I have shown that it is as legitimate as taking a walk then my work is done. A convincing argument that introspection is positively good for AI is supplied in chapter 5, so we could leave it at that here.

However, note the three examples above: Newton's falling apple, Kekulé's dream of a snake eating its own tail, and taking a walk for AI. These examples, though all fine in principle within the context of discovery, are quite different. Newton was exploring *gravity*, and so a falling apple is directly relevant to the content of his research. One could speculate that he was considering the motion of the planets when he saw the apple drop, and therefore could put the two together and come up with his unifying theory of gravity, valid both in the sky and on earth. This example is one where the idea came from something *relevant* - gravity. Kekulé's dream of a snake eating its tail hinted towards *circularity*. Before Kekulé the possibility of molecules' overall structure being circular was not considered, and hence the Benzene molecule was a conundrum. So the dream of the snake gave him the *form* of the molecule. Taking a walk has little to do with AI as such. But introspection, as we will see in chapter 5 and specifically in section 5.2 is of great relevance to AI. See section 4.6 for an answer to any reader that protests that the above is trivial, and already understood throughout the AI community.

4.2.5 Truth in science vs in technology

The focus of the literature condemning subjectivity and introspection (foremost J. B. Watson (1913, 1920)) is in the *science* of psychology, i.e. the development and assessment of knowledge and models about the natural facts about human behaviour. The case of AI as a *technology* is different (Franssen et al., 2013; Simon, 1981), in that the ultimate criteria for the finished AI machine is not "is it true?" or "does it give good predictions?" but "does it work?", "is it useful?", or even ultimately "does the product sell?" (see section 1.3.2).

In science there is great concern for keeping inaccurate or wrong "facts" out of the body

of accepted knowledge. This has several motivations:

- A mistake would produce wrong predictions.
- A contradiction may imply anything.
- In principle, It can be very difficult *logically* to pinpoint any problem once it has entered “the body of knowledge”. For an extreme example see the notion of logical holism in (Quine, 1976).
- In practice, It can take a very long time to locate and root out any mistaken “fact” – for example, the geocentric view in astronomy held for well over a millennium (P. Watson, 2006)[12]. The story of overcoming this model plays an important part in teaching science (Matthews, 1994, p. 165), and hence forms a primary source of science's fear of errors.

In technology the worry about “wrong facts” is much less pressing. If we have a wrong, misguided or inaccurate assumption or model, then the products using the technology would likely not work, or at least be worse than the alternatives, and would be abandoned within weeks (if not minutes), rather than years, decades or centuries. This is based on technology having a short life-cycle, being pragmatic and being *designed*.

When we are faced with a natural phenomenon that we do not understand, our ability to investigate it is limited by every conceivable complication: the system may have subtle interactions within it, may not be isolated from the rest of the world (see for example (Reutlinger, Schurz, & Hūttemann, 2014)), and may have yet undiscovered physics operating in it. Moreover, in science the unexplained phenomenon itself may be holistic - unyielding to our (usually modular) analysis. On the other hand, in technology, we have an *intended* design, often modular, with a well-understood *intended* chain of causation. In the event of a problem we can therefore at least localise which link broke in this *intended* chain of causation. That does not mean that any problem in technology is easily fixed, just that it is easier to recognise a malfunction, localise it, and isolate the “suspect” module, assuming (again) a modular structure with a clear intended causal scheme.

12 Don't confuse Peter Watson, a living Cambridge historian, with John B Watson, the behaviourist. A third Watson will have some comments to make about case based reasoning, in section 7.2. Moreover IBM have a system named “Watson”.

In highly complex systems, like a fully-loaded modern personal computer, there could be tens or hundreds of millions of lines of source-code involved, and no one person, nor even a team that can be gathered in a room can understand what is going on. Moreover, some software is so old that no living person understands how it works. This complexity can create a *seeming* mystery, in that the intended chains of causation (that may malfunction) are not even known, in practise. This does not detract from the fact that in principle, technology can be easily debugged. The difference between principle and practise can involve thousands of programmer-years, but it is still only a management decision to devote the resources to truly debug a system.

A possible retort is that any real artefact outstrips its design, and is a phenomenon in its own right, that may present strange effects in as many ways as a natural phenomenon (an example would be Radium's impact on photographic plates (Mould, 1998)). There are two possible answers: first one could accept the criticism and appeal to common practise, which shows that technological artefacts developing truly unexplained behaviour are much rarer than in the case of natural phenomena. A second reply (and a decisive one in this case) is that in the topic under discussion here we have ideas (of variable quality) eventually to be expressed as software. Software is run using computers, a well-established technology that is specifically designed to behave digitally and deterministically (B. C. Smith, 2005). And so, as long as the hardware platform does not seriously malfunction, the behaviour of software is deterministic and clear, or at least clear-in-principle, as discussed above.

Moreover, if a technology works, the typical attitude is “who cares” about any inherent truth values. Utility is the be-all and end-all of technology. A further point regarding technology with an “interesting” malfunction is that such a technology turns the conversation over to science. An example of both points is the placebo effect. Placebo drugs “shouldn't” work, but they do. Insofar as a medical practitioner is acting as an engineer fixing patients, they will use placebo treatments pragmatically. In parallel, the placebo effect is a vibrant area of research in science. It seems to be in the essence of the engineering professions to not be interested in the seriously difficult issues – leaving these to science[13]. Another example is that animal cloning is often successful, and can be seen as a technology. However, often the cloned animals described in the literature

13 Specific thanks are due to Joshua Weinstein for discussions leading to these ideas.

did not have the same longevity as naturally-bred animals. This anomaly became a subject of research in science (Klotzko, 2001).

So our anxiety to preserve truth from being polluted by falsehood should be much smaller in technology than in science (see also section 1.3.2).

4.2.6 Example & summary of "introspection is forbidden"

An example (for sections 4.2.4 and 4.2.5) of how truth or lack thereof worked out well specifically in AI is the impact of the paper "*A logical calculus of the of the ideas immanent in nervous activity*" (McCulloch & Pitts, 1943). It inspired, over time, the notions of finite automata, integrated logic design (the core of computer electronics) and made a crucial contribution to our AI notion of neural nets (Piccinini, 2004, p. 175). This was all done notwithstanding the fact that the paper assumed "*that mental states can be analysed in terms of mental atoms endowed with propositional content, the psychons, and that the neural correlates of mental phenomena correspond to precise configurations of neuronal pulses: individual pulses correspond to individual psychons, and causal relations among pulses correspond to inferential relations among psychons*" (Piccinini, 2004, p. 205). These assumptions are in retrospect false, but nonetheless caused great advances in AI technology.

Introspection is legitimate as a basis for technology-AI designs because it would be used in the context of discovery, where ideas are born. Any idea so formed would have to be tested empirically later. And even if somehow some piece of "wrongness" creeps from introspection into the final "findings" of the technological research, it would be weeded out quickly because the life-cycle of concepts in technology (especially computer technology) is much shorter than of concepts in science. We definitely do not need "*generally correct* [and] *reliable reports*" for AI. Recall that is specifically what Nisbett & Wilson (1977, p. 233) were complaining about (see section 4.2.2).

4.3 Introspection as "commonplace"

So far I have discussed a mainstream bias of cognitive science – that introspection is "wrong", and hence forbidden. We now turn to two less dominant groups, first, and less distinct, are the "admitters". These people admit to using introspection, notwithstanding

the bias against it (Costall, 2006; J. B. Watson, 1913). Surprisingly, some of the worst detractors of introspection also used introspection (e.g. Herbert Simon, see sections 4.3.3, 2.2); very occasionally they even admitted to it more-or-less openly. In the following section (4.4) I will discuss the introspection-enthusiasts, mainly Dreyfus. The question of whether AI researchers can do anything *other* than introspection will be discussed in sections 4.5, 5.3 and 6.1.4.

This section surveys the evidence both for introspection being used, and for introspection *simultaneously* being frowned upon. This later point is crucial for answering any objection that my thesis is vacuous, in that AI researchers already use introspection freely and fully.

4.3.1 Sweeping testimony

The most sweeping testimony that we have for using introspection are the least specific, and the least personal:

The first and only direct testimony I found from a founding member of the AI community about the overall field of AI is from **Ray Solomonoff** (b. 1926, d. 2009) who was the founder of algorithmic information theory and one of the original participants on 1956 AI conference in Dartmouth (McCorduck, 2004, Chapter 5). He writes (in a sadly neglected paper[14]):

> ***Almost all*** *of the artificial-intelligence work on problems of sufficient complexity...* [that are] *successful frequently enough to warrant trying them... are* ***usually*** *obtained by* ***introspection****; the experimenter is modelling part of* ***his own mind*** *within the machine.* (stress added) (Solomonoff, 1968)

As we will see throughout this section, the testimonies do not specify what kind of introspection was used. This is an indication that introspection was frowned upon in the AI community – for these scholars are not stingy in detail in other matters. For example, we do not see here any explanation about whether the AI work was based on direct perception of mechanism, or was the mechanism inferred? This distinction was discussed in section 3.3.3.2.

Sherry Turkle (b. 1948), a sociologist, relays the testimonies of some principal AI practitioners she interviewed: **Roger Schank** (b. 1946, the father of Case Based

14 To the best of my knowledge this paper was previously quoted only by myself (Freed, 2013) and in a retrospective recounting all of Solomonoff's work (Dowe, 2013)

Reasoning (CBR), see section 7.2) said "*There's only one place to get such ideas about intelligence, and that's from thinking about myself*". And such thinking, we are safe to suppose, was based on self-observation – otherwise why would he be thinking of his *self* rather than some other concrete example. Schank is an extreme example of using introspection but denying it. CBR (to be discussed in detail in section 7.2) seems to be derived from introspection, but I have found not a single mention of introspection in Schank's works discussing AI. In a personal interview with Turkle, he admits what he concealed – see quote above.

Turkle relays also **Donald Norman** (b. 1935) saying "*In the end I have just* [observing] *myself, and if it feels right that's what I have to trust*". **Marvin Minsky** (b. 1927 d. 2016) engaged (amongst many other projects) in building AI for jazz improvisation because he himself was involved with jazz. Minsky explicitly forbade any "psychological data" in his lab - so they used introspection. Minsky explains: "*What you had to do was something like what Freud did. **Tom Evans** and I asked ourselves, in depth, what we did to solve problems like this and that seemed to work out pretty well*" (Turkle, 1984, pp. 265–7).

Turkle explains that this was not seen as the dreaded "introspection" for two reasons: "*First, they say that trying to capture one's thought processes in the form of a program forces you to confront objectively your initial idea of how you think you think. … you can work towards closer and closer approximations of something that will both 'feel right' and 'run' – that will produce the right results*". Second, she suggests that the conceptual world of computer programs somehow provides a better vocabulary than "naïve introspection" for understanding what we see in mental self-observation (Turkle, 1984, p. 267). This idea that one first "captures one's thought" and later experiments with software based on the mechanism of such a thought is the closest precursor I found for my arguments on the contexts of discovery and justification, see sections 4.2 and 4.6.

4.3.2 Specific apparent cases

This section discusses apparent cases of the use of introspection in AI research, but the word "apparent" plays two different roles here: AI researchers admit to using introspection in specific cases – explicitly, and hence "apparently" in the "clear" sense

of that word. However, in other cases the use of introspection can only be inferred, or even speculated, since the prevalent compunctions prevent researchers from admitting to using introspection. These are "apparent" cases in the sense of "probable" or "possible" uses of introspection. The purpose of this section is twofold: to show specific cases where introspection *was* used in AI, and to show how researchers distance themselves from such usages.

A bridging case, both a survey of other people using introspection and a personal specific testimony, is the case of **Phil Agre** (1997, p. 145) who says:

> "*… I began by filling my notebook with exhaustively detailed stories from my own everyday life. By this time I had grown preoccupied with planning research, so I decided to gather some examples of real-life planning. In doing so, I was following an AI tradition of introspection that has been described aptly, if unsympathetically, by Turkle* [quoted above, section 4.3.1]. *Many early AI researchers were clearly attempting, at one level or another, to reproduce their own psyches on computers, and many of them drew on introspection to motivate their programs. Introspection as a formal research method in psychology, of course, had been comprehensively discredited decades earlier. But AI people have not regarded introspection as evidence but as inspiration; because the functionality of their computer systems provides a fully adequate criterion of the success of their research, they believe, it does not matter what experiences might have motivated the systems' design. And introspection is close at hand. But my own practice was different from introspection in one important respect: whereas introspection attempts to observe and describe mental processes under specially controlled conditions, I was trying to remember and recount episodes of concrete activity that took place in my own everyday life*".

Agre's understanding of introspection is quite restricted – he says that it is usually done in controlled conditions – as a psychological technique that was mostly true (in the time before (J. B. Watson, 1913)), but there is nothing in the definitions of introspection as such to require this (see section 3.3.2). Agre's description of introspection as "inspiration" is in a sense a precursor of my analysis in terms of the contexts of discovery vs. justification (see section 4.2.4), and in that sense he is my "nearest neighbour". Agre ended up writing "Pengi" - which was *not* based on his own introspection but rather on Heidegger's philosophy (Dreyfus, 2007). So it seems that Agre is partly confounding three things: 1. His self-reports on his everyday life, 2. Introspection by pre-Watson psychologists, and 3. The tenets of the Heideggerian tradition. Even with Agre (who seems to be my "nearest neighbour") we see an effort to

distance himself from full-blooded introspection - "*my own practice was different from introspection in one important respect...*". This thesis *recommends* introspection as a conscious and unabashed basis for AI development (as detailed in chapter 6). Note that Agre is interested in "planning research" - this is a fairly sophisticated attitude, while my interest here is in uncovering and programming underlying (anthropic) mechanisms rather than the "western, modern, well-trained and adult" or "correct" thinking that most of AI aims at (see section 3.2).

For all his personal research and writing of "detailed stories" of his everyday life, Agre ended up writing his software based on Heideggerian philosophy, even based on a significantly simplified version of it (Dreyfus, 2007). It would not be much of a speculation to say the following: Agre ended up preferring the "respectable" and "correct" scholarly source over the frowned upon or forbidden personal introspection. This preference is unnecessary (in the context of discovery), and is what I refer to when I say that researchers "shy away" from introspection, and refuse to engage with it in a "full blooded" way - and Agre was the bravest researcher that I have found, in terms of explicitly using introspection.

Consider **Alan Turing**'s (b. 1912, d. 1954) work on chess in the 1940s-50s:

> *If I were to sum up the **weakness** of the above system in a few words, I would describe it as a caricature of my own play. It was in fact **based on an introspective analysis** of my thought processes when playing, with considerable simplifications. It makes oversights which are very similar to those which I make myself, and which may in both cases be ascribed to the considerable moves being inappropriately chosen* (emphasis added) (Turing, 1953).

Note that Turing views his using introspection as a basis for a model as a "weakness". It is doubtful whether he is influenced here by Watson, since he was less of a psychologist than other AI developers, but the reservations he had about this usage is clear. Note that he does point out that making similar mistakes would be evidence of being based on a model of a natural mind.

There are two cases where we have some basis to believe that introspection was used though it was not admitted to. The case of **Roger Schank** erasing any mention of

introspection in his research publications while admitting to it in an interview was discussed above (section 4.3.1). The case of fuzzy logic is more speculative:

The field of fuzzy logic originated with **Lotfi Zadeh**'s (b. 1921) paper “fuzzy sets” (1965). This paper gives no hint as to how these ideas originally occurred to Zadeh, other than that he was thinking about how people think rather than computers. The logical possibilities are that Zadeh was considering how concepts, categories, and set membership work for either himself, or other people. The first will be a case of introspection (using his own mind as the basis of his design), while the second would be a case of 3rd person research into how people operate (or speculation). McNeill and Freiberger (1994, p. 15) provide the nearest text we have to a biography of Zadeh, or to a historical account of the “fuzzy” idea. They say

> *... he had promised to work at the RAND Corporation later that month and had not yet chosen a research topic. So he lay down on a bed, his preferred posture for cogitation, and contemplated complex systems. And the notion of fuzzy sets struck him*

Since on that bed in July 1964 (as far as we know) he was not engaged in any interaction with psychological research about humans in general, it would probably not be a great misrepresentation to assume that he was introspecting when he invented the idea of fuzzy sets, and consequentially fuzzy logic (see also sections 7.1, 6.1.4 pt 4).

Note that regardless of whether we accept the speculation that Zadeh was indeed introspecting, this new and unusual notion, based on “how humans think” (subjectively) was immediately couched in respectable mathematical terms, be they “set theory” or “logic”. The pressure of researchers to be “scientific” seems to disallow a full-blooded acceptance of introspection even in the context of discovery. This is a philosophical error, but should not reflect badly on any of the above scholars, as they are not philosophers. It is however the role of philosophy to point out that such shyness in using introspection in the context of discovery is unwarranted and probably hinders exploration of novel ideas (see section 1.3.1).

4.3.3 Mainstream cognitive science uses introspection

Herbert Simon (see section 2.2) continued Watson's (1913) objections to introspection, and took part in bringing much of behaviourism’s heritage into the cognitive fold

(Costall, 2006). However, While developing the Logical Theorist, in RAND at 1955, he testifies: "I *was doing a lot of introspecting on my own problem-solving processes, so I tired to solve some problems from the Principia… I pondered as I walked about how one solves geometry problems… suddenly I had a clear conviction...*" (McCorduck, 2004, pp. 161–162).

Simon, as we saw in section 3.3.3, explicitly pushed for a continuation of Watson's (1920) "thinking aloud" (TA) and wrote extensively about the use of "thinking aloud protocols" (Simon, 1996a), culminating in his book "Protocol Analysis" (Ericsson & Simon, 1993). I have shown (in section 3.3.3) that the distinction between TA (encouraged by the Watson-Simon orthodoxy) and introspection (forbidden by the same orthodoxy) is :

1. The contents of TA is the *contents* of the thinking process. In "classical introspection" the contents of the report are about the **mechanisms** that were used to do the thinking, rather than the **content**.
2. In the forbidden "classical introspection" it is the **psychologist**, a trained individual with a research agenda, who is making the report, while in TA it is a **naïve** participant in an experiment.

Simon used some TA protocols to develop his AI. In developing the General Problem Solver (GPS), he had to bend the rules a bit:

> *But the most massive set of examples of the experimental strategy of 'just looking' is to to found in human problem solving. Density of data was the name of the game, and protocol analysis the way of playing it. Both Al Newell and I agree that the core of GPS was extracted directly from a particular protocol that we can identify. … The GPS theory was extracted by direct induction from the thinking-aloud protocol of a* [single] *laboratory subject, without benefit of an experimental and a control condition* (Simon, 1996a, pp. 384–385)

In using a single protocol from a single subject, he is giving up some of his scientific rectitude. Also, in using a specific protocol he seems to be breaking one of the two distinctions that separate the forbidden introspection from TA - he is indeed using a naïve experimental subject, but he is using a subject not for the *content* of his thinking, but for garnering information about mechanism. This is arguably introspection-by-proxy.

(Nisbett & Wilson, 1977) (though not directly engaged in AI) argue against introspection, and this is one of the most oft-quoted papers in cognitive science (see section 4.2.2). Still, they cannot avoid introspection completely, and admit to it readily. After rounding up much data to show that introspection is wrong, they turn to speculate on the causes of the subjective certainty that people have regarding their own introspections. They present their theory that "*Confidence should be high when the causal candidates are (a) few in number, (b)…. In fact, we appeal to* ***introspection*** *to support this view" (Nisbett & Wilson, 1977, p. 255, emphasis added)*.

4.3.4 Introspection is "commonplace": summary

So regardless of the engrained, sometimes visceral objection to introspection within the AI community (section 4.2), introspection is somehow permissible (at least sometimes), "frequently enough" successful (section 4.3.1), and has been used before in a widespread manner. This statement is not to fuse different thinkers' positions into one alleged chimera-like position, but to point out the accepted overall conduct of the AI research community as such.

Introspection was often tainted (see section 6.3.3) by prejudice in favour of mathematics in that the researchers described their method of solving the problem (for example of chess) not in a neutral, human-like way, but in an idealised manner; as if we humans give equal attention to all areas of the chess board (Turing), and never make mistakes (within the limits of the search-depth - Simon) - as if we were all rational (perhaps even infallible) in all our thoughts.

Moreover, it seems that though many cognitive scientists and AI researchers do introspect, and many even admit to it, they are still shy, timid, and even apologetic about doing so. I argue in this chapter that introspection is permissible (so they can relax). In the next chapter (5) I argue that it is actually a good and plausible basis for anthropic AI (and by implication maybe also for other fields in cognitive science). My problem with all the above scholars using introspection is that they do it too timidly. For a discussion of how introspection could be done more fruitfully see chapter 6.

4.4 Introspection as "desirable"

To recoup: This chapter is about how previous AI scholars related to introspection. Most

scholars objected to introspection, near-universally quoting (J. B. Watson, 1913). I have already (in sections 4.2.4-4.2.5) shown that since we are here dealing with the context of discovery in technology, these objections do not hold. Further I have shown (in section 4.3) that even some of the same people who denigrated introspection used it informally. For completeness sake, I will discuss here Dreyfus's position (Dreyfus, 1979, 2007; Dreyfus & Dreyfus, 1986), who *indirectly* supports introspection enthusiastically (see section 2.3).

4.4.1 Introspection & phenomenology

Phenomenology is a branch of philosophy (written mainly in German) which attempts peer-reviewed reporting on introspection (Gallagher & Zahavi, 2012, pp. 28–29). Some of the main names in this tradition are Husserl, Heidegger, Merleau-Ponty, Gadamer, Habermas and Dreyfus. In AI, Dreyfus (1979, 2007; Dreyfus & Dreyfus, 1986) is a persistent proponent of this view in the field of AI, but he has not made any technological contribution, and hence is considered by many to be outside the field of AI. Wheeler (2005) tries to integrate phenomenology into the mainstream of cognitive science using dynamical systems, action-oriented representations, and other concepts. Winograd & Flores (1986, pp. 27–37) discuss phenomenology under the heading of "Understanding and Being".

Just as the history of AI is intertwined with that of cognitive science (Boden, 2008), so is Dreyfus's critique of AI intertwined with his critique of cognitivism. Note however that Dreyfus does not distinguish human-like AI from the rational type, nor does he distinguish technological from scientific AI (see my distinctions in sections 1.5, 3.2.1, 3.2.4).

4.4.2 The Neisser-Dreyfus debate

Recall the debate between Dreyfus and Neisser, a founder of cognitive psychology, in section 2.1.

On the surface, Dreyfus and Neisser are on opposing sides – the first is an idealist and the second is a reductionist physicalist, the first is interested in subjectivity and the second in objectivity. Arguably they talk completely past each other, and have nothing in common. But they do share this trait: they each argue that *their own* perspective is true, constitutes *reality*, and should be pursued to the exclusion of other positions. This

idea of “one truth” or “reality” that should lord it over all others is dogmatic and unnecessary (see section 1.4.2). It may be the purpose of science to arrive at a singular truth (though neither Popper nor Kuhn nor any other philosopher-of-science I can recall would agree), but it surely is *not* the purpose of technology to arrive at a singular truth (see section 4.2.5). The purpose of technology is to make (and sell) products that people will find useful (and will buy again). So using different perspectives (see section 1.4.2) pragmatically is not only an option but an (technological-economic) imperative.

4.4.3 Introspection vs. phenomenology

The main problem with phenomenology as a basis for AI is that it is very difficult, if not impossible, to write software based on the refined but somewhat vague and sometimes flowery language of the phenomenologists (see Ed Feigenbaum's reaction to phenomenology in (McCorduck, 2004, pp. 229–230)). The phenomenologists are so advanced into researching what it is like to be human, that there is no way to roll back time 100 years or so and ask them to produce simpler models that make sense in terms of data structures and algorithms. Nor would it be fair to impose such a restriction on their field. Hence I suggest introspection (rather than phenomenology) as the basis for human-like AI: introspection by the individual AI practitioners, as a basis for new designs in AI.

There have been several recent attempts to make phenomenology (and introspection) more accurate, more scientific, and more respectable. A good collection of the different approaches is found in (Jack & Roepstorff, 2003, 2004). These approaches, like Hurlburt’s (2011) version of hetero-phenomenology (see section 3.3.1.5) seem to be motivated by making introspection better in the sense of being more accurate, which is commendable, but is in the spirit of science and/or scholarly study rather than in the spirit of developing practical technology. Like Dreyfus’s (1979, 2007) phenomenological “correctness” here trumps technological feasibility. Purity trumps technology – and with all due respect for these efforts they take us further away from implementable technology, so they are currently less relevant to AI.

4.5 Introspection as “unavoidable”

I will only provide here a sketch of an idea, that will be further developed in section 5.3, once some more concepts have been introduced. This idea is novel, and I do not

necessarily hold it as true (this is an open project). All I do here is present this position as another defensible position regarding the role of introspection in AI.

The vast majority of AI, nearly by definition, takes the form of software. And in developing software, one has two options (or a combination of the two), either to use already existing code or algorithms, in which case nothing is novel, or conjuring up something new. And how does a programmer conjure up some new way of achieving a task? This requires them to bear in mind the problem to be solved, and imagine themselves in a sense to be (or be inside) a python interpreter, or an Intel processor, or some such software environment, and ask themselves how they would achieve the task at hand. The code would then be a log of the instructions that the programmer imagined that they would perform in order to carry out the task. So in this sense, all original programming requires one to project oneself, like a stage-actor, into the world of a software environment, and to write a log of all the instructions that one would perform in such a world in order to achieve the task in hand. More detail will be given in section 5.3.2.

If this argument is successful, this considerably weakens any position that seeks to disallow or denigrate introspection in the context of AI as a software-based pursuit. If introspection is inherent to *all* programming, a demand that we write software while not introspecting would be self-contradictory.

A possible counterargument could be that this position regarding the need for introspection for programming is orthogonal to the rest of the discussion about introspection as a source of ideas for even defining the requirements for a programmer. As with the main body of this specific argument, I remain agnostic, and include this point as a contribution to the completeness of the discussion about the legitimacy of a role for introspection in AI.

4.6 A hybrid position

Having discussed all these positions, another possibility arises. Maybe several of these thinkers (without saying so explicitly) hold a position similar to mine: That introspection is legitimate in the context-of-discovery, and that in the context-of-justification it isn't. That could be the reason that several names appear both in section 4.2 (introspection is forbidden) and also in section 4.3 (introspection is commonplace).

This position is speculative, since I have found no mention of these concepts in AI literature. But as I quoted in section 4.3.1, Turkle offers the explanation that using introspection within AI research was was seen as legitimate for two reasons: "*First, they say that trying to capture one's thought processes in the form of a program forces you to confront objectively your initial idea of how you think you think. … you can work towards closer and closer approximations of something that will both 'feel right' and 'run' – that will produce the right results*". Second, she suggests that the conceptual world of computer programs somehow provides a better vocabulary than "naïve introspection" for understanding what we see in mental self-observation (Turkle, 1984, p. 267).

One could now careen into a deeper analysis of which types of scientific method seem to be on each thinker's mind – but that would be highly speculative. The fact remains that no AI researcher has given a justification for accepting introspection-for-ideas while still rejecting it for evidence. We simply have no textual account with which to clarify the situation any further. So let us now assume (in the spirit of the principle of charity) that these AI researchers foresaw my analysis in terms of contexts of discovery and justification, and therefore I have (in this chapter so far) at most reformulated the received wisdom. Unfortunately for my detractors, this analysis cannot stand because if indeed they thought they could use introspection freely in the context of discovery, they would indeed use it *freely*. However, as we have seen (section 4.3), and we will see further in section 6.1, the use of introspection so far has always been minimized, insofar as it existed at all. Zadeh, (see sections 4.3.2, 7.1) for example, comes up with the fuzzy edges of concepts, and then avoids any further "suspect conduct", and retreats to the safety of mathematics. Introspection is always used as sparingly as possible, like a chef using some potent but frowned-upon ingredient: sparingly, timidly, with minimal fanfare – knowing that the dish will be stale without it, but still hiding it as much as possible from the clientèle. One can only speculate as to why this is so. It is no coincidence that Turkle, the sociologist, got nearest to this point: AI researchers view themselves as scientists, and want to be seen as only using good scientific methods. This stands out in Simon's protestations that his deviation from scientific method are OK (see section 4.3.3) and is foretold by Watson's talk of "the scientific man" (J. B. Watson, 1920), see also (Costall, 2006). Recall also Papert's point that we disown our natural

thought processes, and pretend to think logically (section 1.1).

The only example of people using introspection freely are Dreyfus and his followers (ignoring for the moment the difference between individual introspection and systematic phenomenology). But hardly any of these free-introspectors write any code, and the one that does the most of that, Agre, gets condemned for not being "Heideggerian enough" (Dreyfus, 2007) (Why Agre is not celebrated by Dreyfus as a "first step" in the right direction probably has to do with Dreyfus's own derision of the "first-step fallacy" (Bar-Hillel, 2003; Dreyfus, 1979, 2012)). So we have AI researchers that introspect very sparingly, not really capturing much of the benefit of introspection (see chapters 5, 6), and we have serious introspectors/phenomenologists who are "too grand" to write any actual software, resulting in a dialogue of the deaf – few people in the AI community listen to Dreyfus (McCorduck, 2004, Chapter 9).

So the conclusion of the discussion so far is that the liberal use of introspection in the context of discovery should be allowed, but without allowing the introspective process to get out of control, in the sense of becoming either an end-in-itself or simply so refined that programming the models becomes pragmatically impossible. However, some would worry that we still do not know what kind of information introspection would yield.

4.7 Types of truth in introspection

What does introspection give us? It would have been nice if we had direct access to guaranteed truths, or at least clear observations like we have in the objective world. This is not the case. Let us survey what the worries and positions are. In what senses should we expect introspective reports to be true? This survey does not lead to a satisfying conclusion, like other discussions about introspection (see section 3.3.3.4). There is a way forward though, as will be shown in chapter 5.

1. Introspection cannot be scientific. In cognitive science we are steeped in the rationalistic tradition (see section 2.4). One of its main tenets is that we "c*haracterise the situation is terms of identifiable objects with well-defined properties*" (Winograd & Flores, 1986, p. 15). We find this attitude natural in science, technology, law, trade and many other pursuits that are public (external) in nature. However, once we go into the realm of the subjective, there are

precious few “clear and distinct” (Descartes, 1952) identifiable “objects” in our subjective experience. Even the *categories* of “events”, “objects”, “processes”, etc. (that Schwitzgebel (2012) uses in his definition of introspection) do not necessarily apply. Few “events” or “processes” in our mind have a clear beginning, and even fewer have a clear end (When did I *stop* being afraid of X?). **Mental processes are not “moderate-sized specimens of dry-goods”** (Austin & Warnock, 1964, p. 8).

2. Is introspection a good source of information? Introspection can reflect (or not) some objective reality. By many analytical-philosophy positions (influenced by positivism) take reality to be one, *external*, and objective. Introspection does not give us any correct observations, by definition, since it observes things that are neither external nor public nor directly verifiable by others. However, an idealist position such as Berkeley's would say that there is no external reality other than our impressions of such a reality, and that external, objective reality is merely constituted in inter-subjectivity. Even for such an idealist position, though, it is difficult to show how a single person's introspection establishes any *public* “truth” or “correctness”. **Introspection is often wrong about objective matters** (barring solipsism, where there is no objectivity separate from subjectivity).

3. The internal “introspective vision” could see correctly or not what is going on *for us* (bearing in mind that the use of the term “vision” is metaphoric). It is doubtful, therefore, that one can say that there is such a thing as a “wrong vision”. The vision is what it is – it is what was experienced by the individual, either introspecting, or reporting on some events. So we can say that introspective reports are incorrigible in objective terms (Schwitzgebel, 2012). This is not to say infallible - infallible would be about the objective universe, but introspection is about a specific person's subjective world. And how, on what basis, can we question someone else's subjective experience? We have no access to it! **Introspection is incorrigible about subjective matters**. Nothing here guarantees against introspectors falsifying their reports.

4. In describing *any* vision, having some experience or skill in the matter can be of use. Having a broader vocabulary of colours and shapes allows for a better

textual report of a sunset, and there is no reason to believe that introspective verbalisation would not benefit from analogous acquired skills. Also, akin to describing a sunset, once the broader "strokes" of the vision are already described, there is room for more detail, if one has the patience for it. So at this level there are better and worse introspections, though it is impossible to pass external judgement on their contents, and even more difficult to have objective justifications for such judgements. We can assess and appreciate the level of detail as such. **Introspection can often be improved** (see chapter 6). Whether such skill in introspection changes the *experience* itself or just the description is possibly an unsolvable problem. That, again, is a big worry for science but not for technology (see section 3.3.3.5).

5. However, even in seeing a vision, there is a good chance that one's ability to even *see* things (also internally) will be impacted by their **value-systems** (see also discussion in section 3.3.3.4 about the difficulties with neutral observations). This is an internalised (if you will "cognitive" top-down) form of the following concern:

6. In *reporting* the content of one's introspective vision, one might not tell the entire truth, because it may be embarrassing, or show them up as transgressing on some value system. Admitting to having frequent sexual thoughts (Byers, Purdon, & Clark, 1998) or any other socially-objectionable contents may not be advantageous, so reports may be skewed. Also reports may be skewed by theoretical commitments, e.g. a well-trained scientist may claim to be thinking logically and mathematically when that is not actually what is going on. **Introspection can often be polluted by other considerations (s**ee section 6.3.3).

7. But, there is a seeming **contradiction**: How can I simultaneously say that introspection is "incorrigible", and then say it is may be "polluted"? In *principle*, introspection is incorrigible, and anyone engaged in introspection for AI may by all means go ahead and try designs based on their introspection and see what contribution they may make – let a thousand flowers bloom[15]. However, if one

15 I use this expression with great reservations. The original saying, by Mao Tse-Tung, was "let a hundred flowers bloom". It was used in 1956 in his opening speech for a campaign pretending to

were to say that all their thoughts look to them like flamingos, or alternately claim that all their thoughts are mathematical (as Simon and many others in our field seem to do), or claim infallibility, one may reserve the right to have doubts about these introspections and about the utility that AI might garner from them. These doubts are based on the assumption that as humans we share some of our subjective world, similar to sharing walking using two legs. Some people will be different, but not *that* different. But that would just be an opinion.

Gallagher & Zahavi (2012, pp. 28–29) define phenomenology to be introspection refined by the consensus of the phenomenological community. One could wager that introspections that are in line with established phenomenology would fare better than outliers (flamingos) as AI algorithms. The problem we have had so far was not so much with obtaining good introspections or phenomenology (Gadamer, 1979; Heidegger, 1962), but with programming these insights, see Dreyfus's (1979) failure to code *anything* and his continuing complaint that AI is "not Heideggerian *enough*" (Dreyfus, 2007). See section 4.4.3.

But, still some would worry, isn't introspection messy? There are entire lines of argument saying that introspection, as a process, necessarily interferes with the phenomena that is being observed (Schwitzgebel, 2012), that introspection may not be a process of self-observation, but of self definition (Byrne, 2005), And that introspection, like other observations, is theory-laden (Bogen, 2014). Again, if I were seeking to show that introspection is *correct generally* (objectively), then each of these would be a significant blow – and collectively one can see why most researchers simply dismiss introspection, referring to Watson (1913). But here we do not claim *truth*, but ***plausible utility* for AI**. This chapter's aim (and this section's aim in a different way) is only to show the legitimacy, not the plausibility of introspection. Plausibility is the work of chapter 5.

---

canvass the opinions of the Chinese intelligentsia about how the new Chinese state should be run. Whether initially sincere or not, the result of the "hundred flowers campaign" was that once "the snakes were enticed out of their caves" many of these intellectuals were killed or imprisoned. I use this expression here in its innocent connotation (Brown, 2010, pp. 313–318).

4.8 Introspection may legitimately be used for AI: summary

Introspection was considered suspect in philosophy (section 4.1), and forbidden in science (section 4.2). However, in the context of discovery there should be no restriction on what ideas can be considered, and moreover the level of worry about possible errors interposing themselves as truths is lower in technology than in science. So we have seen that introspection is legitimate as a source of ideas, and some developers of AI would (half-heartedly) agree (section 4.3). The problem with the researchers that practise introspection is that they do so half-heartedly, and admit to it quite rarely. In being ashamed of using introspection, they can hardly reap the benefits that introspection may offer fully.

Dreyfus and others considered phenomenology (a close relative of introspection) a promising field of enquiry for AI (section 4.4), but have produced no tangible examples of working designs. I have provided a speculative argument saying that AI *without* introspection is impossible (section 4.5), to be further developed in section 5.3. A further discussion of a hybrid approach, assuming that the AI community already accepts my point about the contexts of discovery/justification produced a recommendation for allowing the *far more liberal* use of introspection in the context of discovery, while restricting any introspection to programmable models.

For anyone worrying about what kinds of truth can be produced by introspection an argument about the logical status of introspection was given section 4.7. This culminated in a dual view of introspection: In principle it is incorrigible, allowing any introspection the chance to be used as a basis for AI designs. In practice, I have conceded that there may be better or worse introspections, and these may have some bearing on the resultant AI. Conveniently in technology there is no harm in letting “a thousand flowers bloom”.

My argument in this thesis is for using introspection wholeheartedly, but without losing sight of the need to produce code (details on how this can be done are in chapter 6).

If introspection is legitimate, we still need to figure out whether we *want* to use it – do we have indications that introspection would be a fecund source of good ideas? This is the topic of Chapter 5.

5 Introspection is likely to be profitable

Table of Contents

5 Introspection is likely to be profitable....122
5.1 Conceptual arguments....123
5.2 An argument from education....124
5.2.1 Skill questions....124
5.2.2 Teaching skills....126
5.2.3 Self-observations....127
5.2.4 Mental self-observation is introspection....128
5.2.5 Examples of mental skills being transmitted by introspection....129
5.2.6 Skills only part-acquired by explicit instruction....131
5.2.7 An argument from education: summary....132
5.3 Programming impossible without introspection....133
5.3.1 Role-playing....133
5.3.2 Programming is introspective....134
5.4 Introspection is likely to be profitable: summary....136

The thesis is that "**introspection is recommended for anthropic AI**". After various introductions and overviews in previous chapters, in the last chapter I showed that introspection is a legitimate method for inventing AI systems. But being legitimate and permissible is not enough for being recommended: this chapter will argue why one positively should expect introspection to be a good basis for anthropic AI. Recall that recommendation is not a guarantee, what I need to show here is that introspection will plausibly be a good basis for AI inventions.

This chapter will include the following arguments:

1. Consciousness is inherent to all normally-intelligent humans, and forgoing an examination of consciousness in the quest for AI would be as odd as forgoing the examination of horse-carriages when trying to develop the horse-less carriage (section 5.1).
2. If teachers use introspection to teach mental skills to their students, then the very survival of civilisations over multiple generations is testimony to the validity of introspection (section 5.2).

3. A case can be made that all software programming is introspection-based and therefore we have already plenty of evidence that introspection is often useful (section 5.3).

The next chapter will go into more detail on how to develop such AI (and which pitfalls to avoid), and chapter 7 will give some concrete working examples.

5.1 Conceptual arguments

One could present several a-priori arguments for introspection-for-AI-design, but not everyone would be convinced by purely conceptual arguments (software engineers are notoriously suspicious, even of their own code). Moreover, one of the most convincing conceptual arguments I have in this context would require assuming idealism, which many would refuse. Instead I will let the next section (5.2) do the heavy lifting of giving a compelling argument. Here, in the conceptual stage, I will only make one argument, and move on.

The main model we have for intelligence is natural human intelligence. This is by definition the intelligence we want to emulate in *human*-like AI. Specifically in this thesis we are looking for anthropic AI, that is the un-enculturated intelligence, prior to any moulding of a particular mind to a particular environment or society. This is contrasted with the "western, modern, well-trained and adult" mind that so much of existing AI aims at. *Artificial* intelligence is an attempt to *functionally duplicate* intelligence, like artificial leather is an attempt to functionally duplicate leather, or an artificial limb is an attempt to functionally duplicate a limb. One of the most basic things we expect an engineer to do when developing an **artificial X** is to examine the *natural* X as much as is needed, to glean ideas for their artefact from the natural X.

Every human mind is subjective, and hence subjectivity is a matter of natural fact (Seth, 2010) inherent to the human mind – the main specimen of intelligence we have available for examination. But most AI practitioners would have us construct intelligence *without subjectivity*. We have no natural example of such an intelligence, and so building AI without subjectivity would be building artificial X while excluding in-principle one of X's most salient characteristics – indeed a bit rather like climbing a tree trying to get to the moon, as Dreyfus (1979) had it. I suspect this is the reason why AI has become moribund in the last few decades - why the initial optimism led to such

meagre results (Langley, 2006; McHugh & Minsky, 2003).

5.2 An argument from education[16]

In this section I will discuss a specific, rather common type of education – the conscious teaching and learning of skills. This is not to detract from all other types of education or other aspects of human existence and development, such as emotions, mood, motivation, and countless others.

I will here argue that when *teaching skills*, the teacher needs to self-observe in order to know how the teacher himself does X, so that he can teach how to do X. In the case where X is a mental skill, then mental self-observation is required. Mental self-observation is introspection. The success human cultures have in transferring their mental skills from one generation to the next is testimony that introspection is neither noise nor nonsense, but is sufficiently useful to allow the new generation to acquire the skills that are characteristic of that culture. If the words used by teachers who introspected are sufficient for the young to become skilful, and people do not acquire said skills spontaneously, it is safe to assume that the words derived from the introspection efficaciously contain the necessary and sufficient information for reproducing the skill (in a healthy human). If the necessary and sufficient information for human skill acquisition is in the introspective text, then: a. Introspection is neither noise nor nonsense, and b. Introspection is a plausible source of information about the skill under discussion. The rest of this section details this argument.

5.2.1 Skill questions

This argument focuses on teaching skills, rather than formal knowledge – on knowing-how rather than knowing-that. This is partially motivated by the conclusion of section 3.2.7 that we should look more at knowledge-how than at knowledge-that. Partially it is simply the best field in which to demonstrate that introspection contains useful information. Even if this argument has no validity outside the “teaching skills” arena, the conclusion that at least in this arena introspection is positively useful – this finding shifts the burden of argument to the other side who might argue that introspection is restricted *only* to that field. Note from section 3.2.7 that the radical anti-intellectual

16 This argument was considerably sharpened by separate discussions with Joshua Weinstein and Simon McGreggor.

position says that there is no knowledge *other* that skills.

-

Having made the knowledge-that/knowledge-how distinction (in section 3.2.7), let's examine how we communicate such knowledge, specifically how we answer questions about such knowledge. The issue of how we answer questions of knowledge-that are pretty well explored in AI (e.g. SHRLDU (Winograd, 1971)). We also have a raging theoretical debate on whether there are any representations involved in knowledge-that (Dreyfus, 2007; Shanon, 2008; Wheeler, 2005). The issue of knowledge-how, and how we answer questions about it requires more discussion.

Note that when one asks a person a seemingly straightforward question such as “Do you know how to ride a bike?”, there are at three different conversations that may ensue:

1. Most people hear it as: “Can you ride a bike”, which would mean “Do you believe that you could ride a bike at will”.
2. A teacher may also hear this same string of words as “Can you teach how to ride a bike?”, and that in turn can be construed either as a request for a yes/no answer, or as a request for teaching.
3. A scientist (perhaps a caricature of a scientific psychologist) may hear it as: “Do you know, in detail, what strategies humans deploy to ride bicycles?”.

These are very different question, but the answer to 3 includes an answer to 2. Consider a slightly different how-to question: “How can I get to London by train?”. Most people, including teachers, would see that as a request for directions, and tell you which turns to take to walk to a railway station, and suggest a train line from the selection available at that station, perhaps with some changes. The caricature-scientist's version of this question would be much longer, but will include also the teacher's version. When asking how humans would navigate to London using trains, after answering all the scientific questions, ultimately the scientist will also need information about station locations, train routes, and schedules. So *both* answers will include train stations, routes and timetables.

In a sense, in order to maintain scientific rigour, scientists *set aside* what they already know as they step into their scientific role (see section 5.3.1 about roles). This is

because scientists see their role as asking the fundamental questions, and demanding detailed objective answers. For technology, perhaps that is a less-than-optimal attitude: a case of allowing the better, more detailed information to get in the way of the good, workable technique. In technology, the every-day human way of doing something is a good first stab at how to accomplish the task at hand. How to fill the gaps between the coarse details of the simpler explanation will be discussed in section 6.3.5.1. Scientists want a whole and fully objective truth, technologists just want something that will work better than the previous technology (see section 1.3.2): Wheels are far superior to beasts of burden in terms of carrying things around, but the idea of using beasts or burden was a huge technological evolution for humans over carrying loads themselves. Everyday people, and technologists, want to get the job done, not to get it done in the best possible way.

5.2.2 Teaching skills

I have found three ways to teach skills (there may be more): explicit instruction, imitation, and “reverse imitation”. One can learn skills by oneself (see section 5.2.6), but here we are discussing *teaching*. These learning methods can be combined:

1. Instruction is when a teacher explicitly uses language to describe how to do something. A good example is a recipe.

2. Imitation is when the teacher provides a demonstration for the student to copy, by observation. Copying by observation may require repeated experimentation by the student, and possibly repeated demonstrations by the teacher. This need not be formally orchestrated – learning by imitation (especially children from parents) happens in the everyday course of life. When a chef puts a video clip on the internet showing how they make some dish, this allows the viewer/student to imitate their method in more detail than a shorthand recipe. Watching the video even without the recipe (perhaps more times) would also allow a willing student to learn enough about how to prepare the dish to write out a recipe themselves.

3. “Reverse imitation” is when the teacher (say) stands behind a student, holds (say) their arms and shows them the correct motion *using the student's own body*. This is done in training beginners in crafts, music playing and sports such as tennis. In a sense the teacher sculpts a moving functioning skilful person out

of the body of a novice.

The first method, “instruction” is fully explicit in that the skills are transferred using language, while the latter methods are at least partially implicit. However, one may object that the textual needs interpretation, while the embodied demonstration is the *more* explicit. This disagreement revolves around the exact definition of “explicit”, but makes no difference to the main argument here.

Not all skills can be acquired by explicit-language teaching. For example control of one's own limbs is developed by trail and error, in infancy (O’Regan, 2011). Even explicit instructions like “add a tablespoon of oil” depend on these basic motor skills which can not be taught. The AI implications of this point will be discussed in section 6.3.5.1.

Fascinating as some of these aspects are, we must concentrate on explicit-language instruction, since we will eventually need to write very explicit AI software. We must focus on how teaching begins, in the case of the explicit use of language in teaching. Our first concern is how the instructions come into existence, in the teachers.

5.2.3 Self-observations

In the case of explicit instruction by words (the recipe of point 1 in section 5.2.2), the teacher has to utter (or write) some text that will communicate the skill to be learned. Where could this text come from? Maybe the teacher gets the exact text from a book – but then the question is begged, how does the author of the book come up with the text for teaching skill X?

1. The text could be remembered verbatim from the time when the teacher/author herself learned the skill. This is unlikely for two reasons: first it would require word-perfect memory over decades. Second this would lead to infinite regress, the instructions had to come from *somewhere*, and unless we believe something like a God giving all instructions to mankind at some stage (on mount Sinai?) we have to accept that the instructions on how to perform skill X have come from some human developing or “coming up with” them.
2. The teacher/author may be flailing in the dark, coming up with various nonsensical instructions, and testing which of them work, and use those. This

idea is a bit like Darwinian evolution. This is unlikely in that we see very few cases of nonsensical instructions being "tried out".

3. The teacher/author may be trying out doing X for themselves (either actually, or in their imagination), while self-observing to see what it is that they are doing. There is some evidence for this in the fact that a student would often ask their teacher (as in the examples above) something like "How would *you* do X"? "You" here is the teacher. The student asks the teacher to observe themselves and tell how X is done. Doing X while observing oneself would also include imagining doing X, and imagining doing X using various options, or recalling a procedure, "playing it our in one's mind", observing *that* and verbalizing it, etc. As long as there is an action by the teacher being observed and verbalised, real, remembered or imagined[17], there is some version of self-observation going on.

A simple example of self-observation would be, when teaching how to ride a bike, the teacher self-observes that in order to get on the bike and start pedalling, he needs to tilt the bike towards his own body so that his first push down on the far pedal will not cause the bike to fall over to the far side. The teacher will also observe what his habits are in other areas, such as where he aims to move the bulk of his body, what angle he points the handlebar at, etc. All these are observations of the preparation to mount the bike – and the more of them can be made explicit the more of an advantage the teacher can give their student. Learning to ride a bike can be done by imitation alone – but with instructions it is learned more quickly (and less painfully).

5.2.4 Mental self-observation is introspection

Some of the skills civilisations possess and pass from generation to generation are mental skills – skills like using nouns to refer to things, skills like deciphering alphabets (as you are doing now, we usually call it "reading") or deciphering pictograms, doing mental maths, and many more. As we have seen above, teaching a skill X requires the teacher to self-observe and communicate how the teacher herself does skill X. If this skill is purely mental, in that there is no external behaviour that can be observed, there are two important consequences:

17 The "simulation theory of cognition" holds that the same mechanisms are used to perform, imagine preforming, and recalling performing any interaction with the environment. This theory is backed with neuroscience data (Hesslow, 2012).

1. Any option of the learning happening by demonstration-and-imitation becomes unlikely, as there is little to observe. A possible exception to this is the bootstrapping of very basic skills like using one's limbs (O'Regan, 2011) of bootstrapping language, a special skill according to Chomsky (Cowie, 2010).
2. The teacher also has nothing external to observe in themselves when they perform a mental skill (no "tilting the bike").

To teach a mental skill, the teacher must therefore glean the information about how they themselves do X by *mental* self-observation. Now recall the definitions of introspection:

> *Introspection refers to an observation and, sometimes, a description of the contents of one's own consciousness* (Overgaard, 2008).
>
> *Introspection, as the term is used in contemporary philosophy of mind, is a means of learning about one's own currently ongoing, or perhaps very recently past, mental states or processes (Schwitzgebel, 2014)*.

Recall also Schwitzgebel's 6 characteristics of introspection, in section 3.3.2.

One may object that the teacher may be *imagining themselves* or *recalling* doing the task and not *actually* doing it. So they are not actually engaged in mental-self-observation (introspection) in real-time, therefore violating Schwitzgebel's condition No. 3, "temporal proximity". How to tell the difference between "imagining oneself", "recalling", and actually exercising a mental skill is hard to define, even using neuro-imaging (Hesslow, 2012). However, these difficulties make little difference to my main point here, that the only way one can teach a mental skill to others is by looking at one's own practice – regardless of the degree to which Schwitzgebel's condition No. 3 of "temporal proximity" is fulfilled, and regardless of whether the skill is being exercised "for real" rather than imagined.

5.2.5 Examples of mental skills being transmitted by introspection

1. Consider mnemonic tricks for remembering words in a foreign language: In Hebrew, the word for house is "*Bayit*". So people like Breuer & Shavit (2014) invent tricks like the sentence "*What a lovely **house**, I think I will **buy it.***" ("*Buy it*" sounds like "*Bayit*") to help people remember the word. How do they know that they have a good sentence if not by playing it out "in their mind" and seeing (introspecting!) if it "works"?

2. Consider the game of “mental chess”. It is like normal chess, only there is no physical chess-set and one has to memorise the board and tell the opponent one's moves using words. This game is quite difficult, and one may come up with some tricks, such as clustering using various groupings, that help one remember the board positions. Again, to share such a trick with another person, the inventor of the trick must introspect.

3. Consider the case of remembering sequences, such as telephone numbers. We know that short term memory can handle only 7 ± 2 units (Miller, 1956). However people can remember significantly longer sequences, using “chunking”. For example, people could spell very long words well before scientific psychology started. Whoever came up with the mnemonic device of chunking saw that it worked in their own mind *before* they explained it to others. Similarly other mnemonic tricks were invented inside people’s minds first and then explained publicly using words. Consider the idea of being introduced to a new person called “Ben”, and trying to vividly imagine them enjoying themselves with all the other “Ben”s that one knows, in the hope that that image will stick in one’s mind, and help in the recall of the new acquaintance’s name.

4. Consider this commonplace trick for handling anxiety: one should imagine the worst possible scenario, and imagine how one would cope in such an eventuality. Having sketched out how one would deal with the worst eventuality, the sense of danger and hence the anxiety subside.

A possible objection here could be that these are all cultural artefacts, and we are looking for anthropic AI, which we defined as being the substructure that allows for culture without including any of the contingent culture. There are two possible responses, one conservative and one more daring:

1. The conservative response is that this observation is correct, but what is shown here is that introspection carries useful information about our mental processes overall. The question of how to distil the un-enculturated anthropic layer is left to chapter 6, with concrete examples given in chapter 7. The short version of how this will be overcome is that we should *aim* in introspection for as low a level as possible, and that we should use introspection iteratively so as to refine

the models used for AI. Recall we are not doing science here, and if we do not get something 100% correct that of not a catastrophe – technology is be far more tolerant of errors that science is (see sections 1.3.2, 1.4).

2. The more daring response adds that mental skills, as exposed by introspection, are *very* near the anthropic line. Consider that the examples above contain not only very high level[18] skills, like "mental chess" but also some low level skills like dealing with anxiety – an emotion (example 4 above). Anthropic AI was defined as aiming at the innate mechanisms that enables our culture without being part of any contingent culture. So looking at basic skills, as basic and infantile as possible, brings us near the elusive boundary between the cultured and the pre-cultured. We may venture to use some implications and speculations to reach an even closer approximation (recall section 3.3.4) – let a thousand flowers bloom. Recall again that we needn't reach the "true" "precise" boundary, since we are not looking for scientific or philosophical absolute truth, we are looking for a technological approximation.

More discussion of this point is found in section 6.3.4.

5.2.6 Skills only part-acquired by explicit instruction

If one were to examine skill acquisition more closely, as did Dreyfus & Dreyfus (1986), one sees that there are stages in the acquisition of skills, including for example reading (for the current purpose I will limit the discussion to English). In explicit teaching by an instructor, only a very basic and slow level of the skill is picked up. A simplified communicable "base" skill (involving deciphering every letter individually) is explicitly taught, with the intent of the student eventually discovering for themselves the "higher" variant of the skill which is superior, but difficult to communicate, perhaps partially because of a limited vocabulary, perhaps because we need some experience with the "base" skill before we can discover the "higher" skill. Only after a lot of experience do we have the fluency that allows us to read entire words (sometimes even while not noticing systematically jumbled text (Rayner, White, Johnson, & Liversedge, 2006)).

Does this have any effect on our argument on AI?

18 Note the use of "high level" and "low level" to describe different things that humans do. This tradition goes back as far as Plato, but has shaky foundations. I use it here in the accepted "intuitive" sense, with reservations.

First, the main thrust of section 5.2 is that skills are taught by teachers practising self-observation in order to generate their speech. At no time did I claim that *all* skills are taught. All that was argued is that *some* skills are taught, and that this teaching requires self-observation on the part of the teacher, and mental self-obseravtion is introspection, so introspection is key to the transfer of mental skills between generations.

Second, this issue demonstrates that some skills are acquired without instruction, and therefore maybe the "true" key for learning human skills is *not* introspective. This possible critique misses the mark in terms of this thesis in that "introspection is recommended for developing anthropic AI" does not claim that introspection will be key to *all* future AI development – that would be the sort of "one truth" dogmatism diametrically opposed to the perspectivist attitude of this thesis, see section 1.4.

5.2.7 An argument from education: summary

If person A invents some mental skill and they teach it to person B, they must be using introspection in order to describe what it is they are doing internally. They cannot directly describe neural or cognitive processes that are not conscious, since they not only do not know their own "scientific" states, they also cannot communicate to the student which correct "scientific" processes the student must adopt (their brain or "metal scape" might be different). Later when person B teaches person C, and C teaches D etc., they all use introspection to tell each other how the skill is performed. The case is even stronger in that even if people teach what they are taught using verbatim memory, they must use introspection to describe their own innovations and improvements of mental skills.

These skills (that *work*, somehow, in the real world, using wet neurons) are transferred (and improved) from one generation to the next, by little other than introspective-based narrative.

In teaching, the text "take that digit, and then add it to the other one" is a description of what the teacher thinks (or introspects) she is doing. It can't be mere repetition because that would not account for innovation. If introspection were noise, no student could pick up a skill by this normal method of teaching, but we all do, ergo introspection may be

inaccurate, may not descriptively reflect neural reality, but still **reflects enough of *some* reality** to allow the transfer of skills and culture over many generations (reliably enough for us to read texts from millennia ago).

So (at least some forms of) the contents of introspection have some validity, in a pragmatic ("it works") sense.

5.3 Programming impossible without introspection

Although this idea has already been sketched out in section 4.5, here I will discuss it in more detail in order to show that introspection is not only plausible and recommended for AI, but arguably necessary for *any* programming. But there is a pitfall: one might argue that if indeed all software development is introspective, what is the novelty in this entire thesis? I will discuss this towards the end of this section.

5.3.1 Role-playing

When acting and communicating, humans assume a certain role (or "frame of mind"), usually depending on the social context. The classic example is the way people take on their "work persona" as they start their work day. This "work persona" is in the main similar to the job description the organization would advertise to fill that role if it were to fall vacant. As Herbert Simon (!) noted:

> *Administration is not unlike play-acting. The task of the good actor is to know and play his role, although different roles may differ greatly in content. The effectiveness of the performance will depend on the effectiveness of the play and the effectiveness in which it is played. The effectiveness of the administrative process will vary with the effectiveness of the organisation and the effectiveness with which its members play their parts.* (Simon, 1976, p. 252, 1996b, p. xii)

Another example of people semi-automatically fulfilling a socially-constructed role is that a bilingual person will usually speak only one language at a time.

There seems to be no such thing as the mind operating (in a way that could be relevant for action) outside of some cultural context (see Wittgenstein, 2001a) even if it is the context of running amok (Carr, 1985). This fact of the individual's behaviour being constructed in (usually) socially-accepted roles is transparent to us in daily life, but has been the subject of much research, see for example the much acclaimed book "The presentation of the self in everyday life" (Goffman, 1971).

One of the roles one can adopt is the role of being cooperative with some scientific programme, such as Watson's (and later Simon's) “thinking aloud” (see section 3.3.3.1), requiring that a “*scientific man*” take on such a role as thinking aloud “*in the proper spirit*” and possibly even “*with zest*” (J. B. Watson, 1920, p. 91).

These roles that we adopt come with certain prejudices in interpreting our environment. Again to take an example from bilingual people, the same utterance (the same series of phonemes) may be interpreted completely differently in the context of one language or another: “me” in English is the first-person-accusative pronoun, in Hebrew the same sound (מי) is the personal-interrogative, meaning “who?”. Also, the same event (say the firing of a pistol by an assassin) could be interpreted differently by the same person, depending on whether they are acting in their capacity as a citizen of a polity, or in their capacity as a scientist. In one case one would present the events as an assassination, and in the other one would explain the chemistry of gunpowder and the mechanics of the revolver.

This observation (that humans usually act within a context of a role) is not entirely alien to the field of AI. Note that the above quote is from Herbert Simon (albeit from his work in public administration, not AI).

5.3.2 Programming is introspective

This is a speculative position, which I here only claim is *defensible*, not necessarily true. Arguing properly for this position is outside the scope of the current project, and the main thrust of the thesis does not depend on it.

In writing new code (not debugging or reusing existing code), in a sense a programmer projects herself (as an actor would project himself into a character see section 5.3.1 above) into an imaginary world where she is (say) inside a world consisting of the python instruction set, or in a world comprised of an “Intel” architecture, and asks herself how she could use the tools available (variables, arrays, loops, libraries, etc.) in order to achieve a task such as calculating VAT or whatever the programming task is. There is a lot of “first person thinking” going on, as in “how could I do this”, “this could give me *that*” etc. The programmer's output, the code that is supposed to do the task, is a formalization into python (or “Intel”) *instructions* by the programmer introspecting inside this “world of python”. Where else could the code come from? I

can find no evidence (or testimony) that there is anything like a tree-search of possibilities as GOFAI would have it. Note also the language that is used for the text that the programmer writes in order to invoke a feature of python: it is a “command” or “instruction”. These words, outside the context of computers are used in education in management, telling people *how* to do some bigger task by giving them details of the component-tasks in simple language (see section 5.2).

Moreover, in debugging, a similar thing happens. In the exercise known as “a dry run”, the programmer projects herself, like an actor, into the role of a python interpreter, and acts (in her mind, perhaps using pencil and paper) on the code and the data as a python interpreter would, always keeping half an eye on the intended result to see where the actual result deviates from the intended result. When such a deviation is found, the programmer would say that she found a bug (recall the “intended chain of causation” in section 1.3.2).

Conversely a programmer copying an algorithm from a book is not introspective.

-

In the beginning of this section (5.3) a worry was presented that if indeed all programming is introspective, than what is the point of this whole thesis? I have three retorts:

1. The distinction is subtle but clear-cut: In programming one projects oneself into a formal system (python/Intel instructions) and tries to achieve a task (say VAT calculation) inside that formal world. Conversely, the method promoted by this thesis to get to human-like AI would require us to observe our thought processes *in natural form*, describe them as best we can (see section 6.2), and only *later* formalize them as code. The difference is whether we do the introspecting inside or outside of the formalized world of programming.
2. Another retort could be that if introspection is indeed so wide-spread and accepted, then why is it so often denied and obfuscated? This may have to do also with the next point.
3. There is a distinction (Chrisley, 2003) between something being ontically novel, as in a novel entity, vs being notionally novel – the phenomena was around for a

while, but we never noticed. As a bare minimum, the idea of introspection for AI is notionally novel.

5.4 Introspection is likely to be profitable: summary

This chapter deals with why one should positively expect introspection to be a good basis for AI development.

Section 5.1 argued conceptually that in developing a artificial X one should not shy away from examining *all* aspects of X – so in developing an *Artificial* Intelligence one should examine all aspect of our natural human intelligence, including the subjective-introspective.

Section 5.2 argued that introspection is neither nonsense nor noise, even using the worse-case delineation (see section 3.3.4). Rather it is used to allow know-how to pass from person to person, inducing across generations. Perhaps this very ability of humans (to introspect) evolved specifically to allow the transfer of know-how from one generation to the next, allowing civilisations to survive by accumulated wisdom.

Section 5.3 tentatively argued that *all* programming is introspective, and showed that that does not trivialise the overall argument. Being tentative, this argument is not necessary for the overall thesis and was included for completeness.

As we have seen there is no point in looking for "truth values" as such in introspection-for-AI. This thesis promotes using introspection as a source of *ideas* for anthropic AI as a technology, and that is justified by three moves:

1. The requirement for "truth" is shown to be secondary in technology to the requirement for utility, and the costs of a factual mistake are shown to be lower in technology than in science, at least as far as software is concerned (see sections 1.3.2, 1.4).
2. The argument from the "context of discovery" allows for *any* source of ideas, thereby making introspection as legitimate as any other source of ideas (see sections 4.2.4, 4.2.5). This may be seen as too weak, as by a similar token one could argue for taking a walk in the park before thinking about AI, but in principle this is enough, because of (*inter alia*) the next point:

3. The argument for the validity of introspection for education (section 5.2) shows that introspection is not only better than noise, but arguably the foundation of many of human culture's successes. It is therefore expected to be a positively *good* source of ideas for AI.

This is also supported historically in that some of the more open practitioners of AI admitted to using introspection as the basis for their AI (see section 4.3), even though they were acutely aware that introspection has a "bad reputation" in psychology. They however never used introspection in the wholehearted manner that is recommended in this thesis.

The argument of this thesis is that AI based on introspection would produce better anthropic AI. There is no hard-and-fast logical guarantee that introspection would always produce a good design, but it produces, as a minimum, "*a sketch of a sketch*" as Simon had it (McCorduck, 2004, p. 246), a basis for an idea for a design (see examples in chapter 7). The resultant designs must later be evaluated experimentally, possibly improving the design by iterative introspection and coding.

Let's look at how this is done in detail.

6 Details and how to use introspection for AI

Table of Contents

6 Details and how to use introspection for AI....138
6.1 Definitions and delineations....139
6.1.1 Definition for “AI based on introspection”....140
6.1.2 Non-human-like inspirations....141
6.1.2.1 Genetic algorithms (x2)....141
6.1.2.2 Neural nets....142
6.1.3 Human-like inspirations (non-Introspective)....142
6.1.4 Types of introspection for AI....143
6.2 The process of introspection for AI....146
6.3 Comments on the process of introspection for AI....148
6.3.1 Introspection is a witness account....148
6.3.2 Looking / listening for....150
6.3.3 Pollution....152
6.3.4 Introspection: is it above or below the culture line?....153
6.3.5 Interpolation and approximation....154
6.3.5.1 The holes in introspection....154
6.3.5.2 Opportunistic approximation....155
6.3.5.3 Analogue cannot arise out of digital....155
6.3.5.4 Being analogue does not mean it is not digital....156
6.3.6 Multiple iterations, multiple mechanisms....156
6.3.7 Personnel....157
6.4 Project expectations....158
6.5 Testing and evaluation....159

This chapter outlines how to approach developing introspection-based AI. It also answers several questions that remain outstanding about the recommended approach, and how to go about doing introspection more “correctly” than some of my predecessors did (see sections 4.3, 6.1.2, 6.1.3). In the chapter and on I use a new notation: <u>underlined text</u> designates introspective reports.

There is a seeming tension in my argument, which is about to get worse. On the one hand I promote a liberal approach to the possible bases of (or inspirations for) designs in AI: “let a thousand flowers bloom” (section 4.7, point 7), while on the other hand this chapter will prescribe “correct” ways to do such introspection. My position is that all

types of introspection are legitimate sources of ideas for AI (chapter 4), and some types are recommended as such sources (chapter 5). In this chapter I again *recommend* certain avenues as being more plausible, more "correct" than others, and I give my arguments. These recommendations here do not detract from the generality of the argument in previous chapters.

There are a few other points that may need re-emphasising:

- Human-like AI is not better than rational AI, it is just different, and what this thesis is about, (see section 3.2.2 for motivations) and hence is our focus here.
- Any inspiration or basis may be used to develop any type of AI. In the extreme, this is a form of the argument that monkeys typing randomly for eternity would produce the entire British Museum's library (Borges, 2001). Therefore, from looking at an algorithm we can never deduce with cast-iron certainty its provenance in terms of inspiration or basis (or the species of the designer). In the rest of this chapter I will say things such as "neural nets are based on biology" as shorthand for something like "it is most probable that the AI method of neural nets was inspired by (a grossly simplified version of) real life, wet neurons".
- It seems that so far the vast majority of AI has been based on either mathematics (broadly construed), biological inspirations, or on cognitive models (see (Langley, 2006)). These approaches have been so successful that often when developing new AI many researchers tended to "look under the lamp" in one of these three areas. This thesis is about broadening the search, and recommending direction(s) to explore within the multitude of neglected possibilities.

6.1 Definitions and delineations

Here I will clarify what is meant by AI "*based on* introspection", and contrast the recommended methodology with other (existing) approaches to AI.

In delineating the different AI efforts so far one should look at the following aspects:

- What type of AI is being aimed at (human / insect / rational).
- What were the principal inspirations or bases for the technology.
- How successful the effort was, empirically.

- Especially in case of failure, does the result call into question the qualities of the *entire approach / inspiration*, or only the circumstances of the specific experiment?

In section 6.1.1 I will set out some definitions as to what precisely is being recommended in this thesis. In later sections I will compare and contrast this with how other approaches seem to work.

In contrasting my view of how to develop AI in detail with previous approaches, this chapter will also answer in more detail than section 4.6 any objection that there is little novelty in this approach to introspection-for-AI. The below survey does not claim to be complete, but to present an extensive sample of how AI systems were based on various models.

6.1.1 Definition for “AI based on introspection”

As the idea of “an AI design based on introspection” is key to this thesis, let's define it and look at some examples that fall outside and inside this definition.

Y is based on X

In this context, Y, a design for an AI system, is based on an observation X (that could be an introspective observation) iff:

A) There is a causal link from X to Y.

B) X is the dominant influence on the workings of Y, i.e. there is no significant pollution by some other factor such as a prior theoretical commitment. In our case, of AI based on introspection, this would require acceptance of introspection (X) as legitimate, not to be obfuscated or denied; minimisation of attachment to or influence of any theoretical framework, such as mathematics, logic, or some theory in cognition, psychology, religion, or even phenomenological literature. See section 6.3.3.

C) Corresponding functions are achieved in similar ways (data flows, data structures, temporal order, etc.). Examples of similarities of process and data flow are given below in sections 6.1.2 to 6.1.4.

Introspection (as a basis for AI) further requires

D) Introspecting, i.e. “looking” at or “listening” to one's own *untrained* mental processes. See section 3.2.4 for an argument for “untrained” thought vs “western, modern, well-trained and adult” thought. See a summary of Schwtzigebel’s (2012) definition of introspection above, section 3.3.2 .

E) Proficiency in conceptualising and expressing the contents of this introspective “vision”, in text, diagrams, algorithms or suchlike. (see sections 5.2, 6.2)

F) Fidelity, that later could be judged (possibly externally) as credibility. Introspection is a type of witness account. The fidelity / credibility of witness accounts in general is discussed in historiography (see section 6.3.1). For example, a design that never makes a mistake is not credible (as a human model), since humans make mistakes. Another example: a design that fits some theory too well might well be a result of theoretical pollution, see B above.

Let’s look at all the ways this can fail, or (just another way of looking at it) how AI development has happened so far, and how these scenarios differ from the above:

6.1.2 Non-human-like inspirations

Some algorithms have clear inspirations, which are *not* human, let alone introspective. There is no specific reason to believe these would be good at producing human-like behaviour.

6.1.2.1 Genetic algorithms (x2)

The idea of the genetic algorithm (**itself**) has a clear inspiration – a (somewhat naïve) scientific model of biological inheritance and evolution (usually in a sexually-reproducing population). This is based on biology (and fulfils A-C), but biology is not introspection (and does not fulfil D-F).

A genetic algorithm can be used to generate an **output**, any information-entity for which there is an “objective function”, and as a specific case a genetic algorithm can be made to evolve software for any purpose implied by the objective function. Such a successfully evolved design, since it was not made by any person, can be said to not be based on anything. That is not to say that it was not based in some way on the objective function (it was), but the loops and variables inherent in the code produced by a genetic algorithm are bereft of any human design or inspiration at all. This obviously does not

fulfil any of A-C or D-F.

6.1.2.2 Neural nets

Similar to the case of the Genetic Algorithm itself, neural nets are based on biology, so likewise it fulfils A-C, only regarding biology, not introspection.

One could argue that introspectively it “feels like” our mind functions in a manner similar to a neural net in that there are alternating phases of chaos and stability, thereby fulfilling the criteria (D-F) for our experience as revealed by introspection, but:

- The neural-net's **data structure and data flows** are not visible to us humans internally, therefore are not introspective, but originate in scientific external observation (also simplification - (McCulloch & Pitts, 1943)) of neurons (not output of introspection) – this violates C.
- The observation that the end result of two processes is similar is not the same as observing the process that produces these end results. Such an **observation** would be introspection, and is lacking here (See more about similar results in section 6.3.5).

Again, Neural Nets are based on biology (and fulfil A-C), but biology is not introspection (and does not fulfil D, E-F are debatable).

6.1.3 Human-like inspirations (non-Introspective)

See also section 3.2.5.

Behaviour: An example of building an AI system or robot that was based solely on externally observable **human behaviour** would be hominid robots in general, and a more specific example would be bi-pedal walking robots. This approach is scientifically unassailable, as it is based on hard-core observable facts (humans have two legs, knees, etc.) and is reminiscent of behaviourism. The disadvantage is that the behaviours produced by following this inspiration are not sophisticated. Any level of sophistication would require having some sort of mental processing (this was Chomsky's cognitivist argument in his critique of Skinner (Chomsky, 1959)). Mental processing breaks the behaviouristic framework, and has to be based on something other than behaviour as such.

This example fulfils A-C for behaviour, not introspection, and therefore does not fulfil D-F.

External analysis of human behaviour: One could create a far more sophisticated *model*, like (in natural language) a grammar, and try to base the machine's behaviour on such a model. This is only as good as the underlying model. Our ability, thus far, to model humans is limited. Alternately, one could have models or theories of how the (maybe even entire) human mind works (cognitive simulation (Sun, 2008)). A-C hold for the abstracted models (specific or general), D-F do not hold.

Substructure: Some would simulate the brain in its entirety, either at cellular or atomic levels. This approach is currently not feasible (see section 6.4). Again, one could have a more abstract theory or how the mind is constituted, like logic, statistical inference, other mathematics, or some form of symbol processing (classic AI). A-C hold for the idealised model (or the entire brain), D-F do not hold. Newell & Simon's (1961a) insistence that their subjects think in "rules" can be construed as an example of this (see section 6.3.3). A more brazen example is (Bringsjord, 2008).

6.1.4 Types of introspection for AI

Within the field of AI, there are several variants of introspection used (moving from the existing to the recommended and beyond):

1. Taking a literal view of sentences of the type to be discussed, and a somewhat idealist metaphysical stance, one could see every conscious human observation as arguably introspective. Saying "*I* see red", or "the measurement shows *me* 97" is introspective. Such **trivial introspection** does not lead to software designs, so there is no "based on" relationship (no A-C).

2. Much of **programming** is introspective: As we saw in section 5.3, arguably all programming is introspective – It requires the programmer to imagine themselves as living in world of (say) python features, and asking themselves "what would I do", etc.

   This introspection while "putting yourself in the computer's shoes" is introspection on *learnt* skills, the skill of thinking like a computer, a fundamental skill of programming. One cannot introspect in this way unless one

knows how to think in terms of code. This specifically violates D's requirement for *untrained* mental processes, and it is not consciously based on introspection at all (A-C, D-F).

The difference between my advocacy for introspection and this model of programming is subtle, yet clear-cut: programmers *first* adopt the technological role of "being a computer" and then ask "how would I solve that if I were a program", while I advocate asking "how do *I, as I am*, solve that", and only later turning that fully-human introspection into technical code (see section 5.3).

3. A more general case is introspecting artefacts of culture – "I think **logically**" "I think using words" or even "I think in Java" (as above). The general case is "I think using P" where P is a construct found in a certain culture. Most often in AI, it would be the epitome of what this thesis is *not* aiming for – the terms of the introspections would be western, modern, well-trained and adult rather than anthropic (see section 3.2).

   Case-Based-Reasoning (see section 7.2) can be seen as an example of AI based on such introspection, in that it operates like a western well-trained administrator, using the best solution for any given problem. Again, this violates D's requirement for introspecting *untrained* mental processes. The problem with trained mental processes is that it gives us information about how *our culture* thinks (or how our culture thinks one *should* think) rather than how we *actually* think.

4. I argue for introspection on *natural* mental processes, not enculturated ones (see section 3.2.4). This can be done minimally, by taking a **particular element** of how we think, such as the fuzzy edges of concepts, and building an AI concept around it. This is good as far as it goes, but often is then used again within some well-understood mathematical scheme, such as fuzzy *logic*. The novel idea here (fuzzy concepts) fulfils A-C and D-F, but it is then embedded in a logical-mathematical framework with all its non-anthropic characteristics (One cheer for fuzzy logic! see section 7.1).

5. I further argue here (in section 6.3.6) for introspecting **multiple novel elements**, for a more complete model of how our subjective mind works. These elements

can be added gradually, interspersed with experimentation for feedback on how each step of the development of the AI design works, and allowing further introspection at each stage, looking at how our own thought-processes differ from the model. See examples in chapter 7.

6. Note that I am *not* aiming for “Heideggerian correctness”. Unlike Dreyfus (1979, 2007), one can be agnostic about the correctness (or usefulness for AI) of any particular volume of phenomenological literature. Unlike Dreyfus I argue that AI practitioners must remain committed to pragmatism, to programmability. So there are two things here: One is that the recommended technique is not married to the idea of “correctness”, “precision” or “truth” in general, and the second is that I remain agnostic specifically as to the truth or lack thereof in the German phenomenological literature.

So subjectivity is not only necessary for any conscious human activity as suggested by point 1 above but is needed for all programming (point 2 above) and has been used by several if not all AI programmers before, see section 4.3. But so far AI researchers have been subjective and did introspection in a bashful “under cover” way. (see sections 4.3, 4.6). As suggested above, it may well be time to to do it consciously and properly, rather than coyly as if hiding from Watson's (1913) or Simon's wrath. We surely cannot hope to excel at developing human-like AI while pretending *not* to be human (see section 5.1), as in cases where the inspiration or basis for the design is non-human or an idealised human.

6.2 The process of introspection for AI

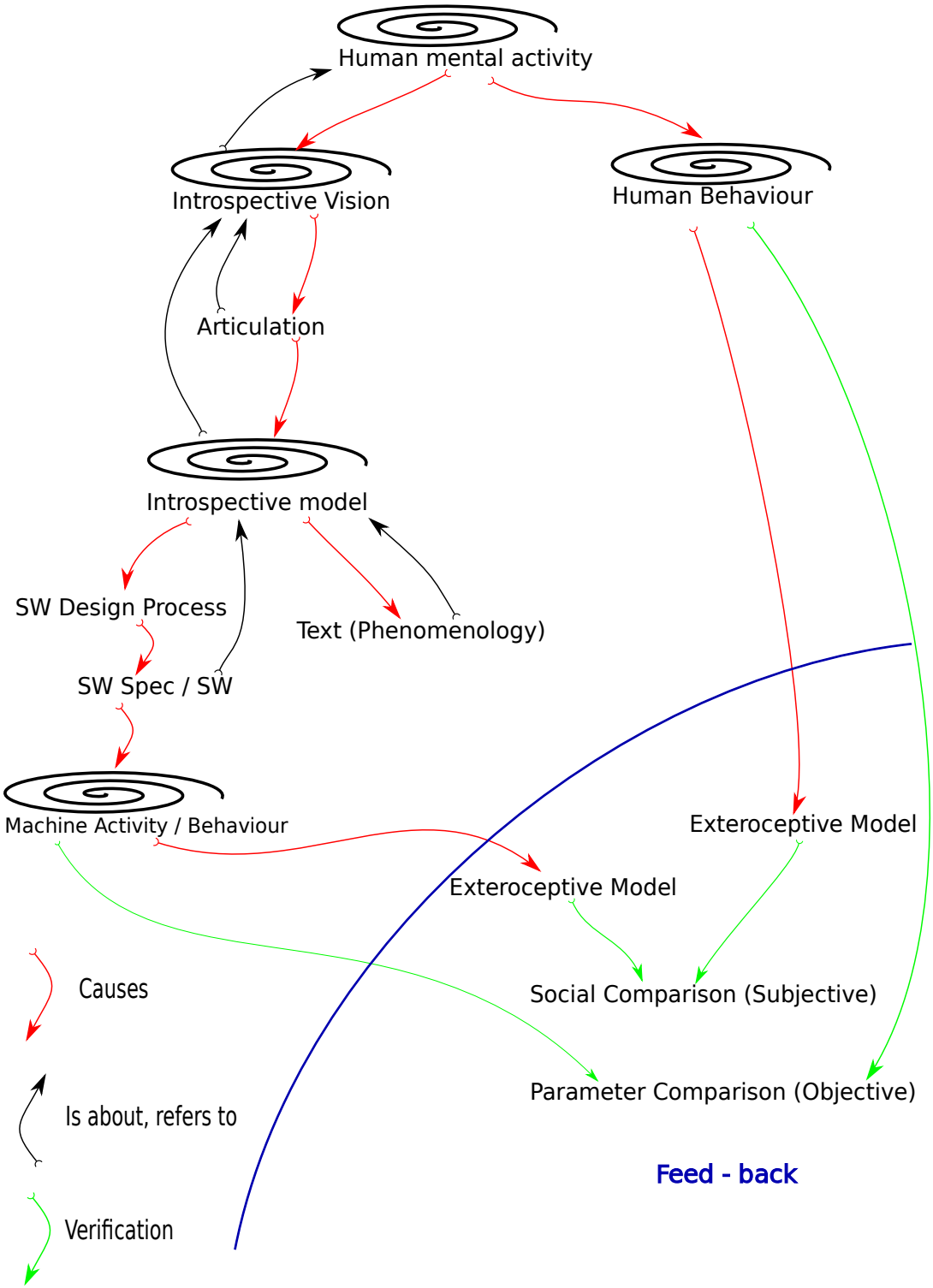


Illustration 6.1: Process of introspection for AI

Consider Illustration 6.1 (bold terms refer to labels in the diagram).

In approaching the process of introspecting for AI one should have an intention to perform the introspection in the spirit of the description to come in section 6.3 (not polluting with theories, “looking for” processes rather than beliefs, etc.). For now lets look at the *structure* of the introspection-for-AI process.

Human mental activity, as seen subjectively (in the phenomenal mind, not the cognitive mind, see section 3.2.5.3) is a process, symbolised here as a spiral, that determines **human behaviour** (red arrows signify causation). **Human mental activity** can be observed by an **introspective vision** (bearing in mind that the use of the term “vision” is metaphoric), which (at least partially, see section 5.2) mirrors the process of the **human mental activity** (hence the spiral here). In this stage, of the “vision”, the issue of “looking for” comes to bear, see section 6.3.2. Next, there is a process of **articulation**, that does *not*, in a sense, mirror the original mental process but rather moves away from it, in forcing the vision into a more communicable form. Here is where “pollution” is a danger, see section 6.3.3. The process of articulation may be iterative, with the vision being further and better understood with additional attempts at articulation. The **introspective vision** refers to (black arrow) the **human mental activity**. **Articulation** produces an **introspective model** – some concrete idea of how the **human mental activity** appears to work. This model can be explained or expressed using **text**, which would be a phenomenological *report* about the **human mental activity**. We should keep this distinct in our mind from the philosophical tradition of phenomenology, which some may argue against, see section 3.3.1.5. Note that a text is just a string of letters – it is not a machine or a mind or a brain that can carry out a process (hence no spiral). The **introspective model** can also be used by a **software (SW) design process**, involving approximating (see section 6.3.5) and digitising, that will produce a **software specification**, and ultimately **software**. This software, when run, creates machine behaviour, which hopefully in some approximate sense reproduces the original **human mental process,** or at least the **introspective vision** (hence the spiral).

The output of this entire process can later (see section 6.5) be compared (for feedback purposes) to actual **human behaviour** in two ways (see area under blue arc). One

alternative is objective **parameter comparison** – observed directly (perhaps even mechanically) in **human behaviour** and the **machine activity**. A second feedback method can involve people observing both the human and machine activity, and producing their **exteroceptive model**s of the two activities, and compare them subjectively. Such a process is similar to (most construals of) the Turing test.

Note that there is no necessity for one person to be involved in the entire process. One person can have the **introspective vision**, do the **articulation,** and express his **introspective model** as **text** (written or spoken), and another person (or several) can engage in the **software design process**, etc.

Note also that the entire process can be an iterative process (not just the articulation), where refinements of AI designs are made by further introspection, modelling, etc. Introspection can be refined in the sense of adding details, sometimes even changing the fundamentals of the picture. For example, a trivial (non AI related) introspection could be "I think it will rain today". A deeper observation of a similar situation can yield something like "By my experience these clouds don't mean much either way, so I can't really say, but I fear being caught out and getting wet, and even more I fear being blamed by my friends for not saying it would rain. So I say it will rain even though I don't know". Notice that the underlying dominant influence has shifted from *thinking* (that it will rain) to being socially *afraid*.

Another way that introspection can be refined is by the development of terminology. Once one is versed in terminology (in any field, I see no reason why introspection should be an exception) one can describe more using less words. This ability to economise and increase precision in language, useful in any case of description, is of particular importance in introspection because introspection happens in the same mind where the phenomena being observed are occurring, and the danger of interference between the two processes is ever-present (and always to a degree unavoidable). Again, Schwitzgebel (2012) gives an overview of this issue.

6.3 Comments on the process of introspection for AI

6.3.1 Introspection is a witness account

When one reports about one's introspection, one gives a testimony about what one had

“seen”. There is every reason to expect that any known issues of reliability in testimony would arise in the case of introspection, so it is worth summarising one of the canonical sources on the technique of writing history, “The historian’s craft” (Bloch, 1953). Bloch starts (pp. 79-81) with observing that believing everything that is presented as evidence is naïve, but also *mis*trusting *all* evidence from a source willy-nilly just because some of the evidence presented by that source is questionable – would also be equally naïve. Often fashionable inaccuracies are accepted as truth, and one must especially be careful with “common sense”, since firstly common sense is nothing other than the fashionable prejudices of a certain society, and moreover common sense changes over time, usually with no fanfare. Often convenient rumours are not denied, because they suit some agenda (*Ibid.* pp. 99-100).

The word “sincerity” has a very broad and dangerous meaning, it can incorporate many of the above prejudices - “*many witnesses deceive themselves in all good faith*” (*Ibid.* p. 100). People are often unaware of the simplest things around them, such as the number of windows in a familiar room. In our case, introspectors may need to practise (*Ibid.* p. 102) in order to see the obvious about their actual thought processes, since they are so familiar (see also (Schwitzgebel, 2004)).

In history, the determination of the exact cause of an event is difficult in the extreme (Bloch, 1953, pp. 103–4), but that is also the case in chemistry: we do not know which molecule in the mixture made the whole thing explode. We can only describe the antecedent conditions, not the exact cause. Also, no witness's testimony is equally reliable on all matters (*Ibid.* p. 101), but in our case of introspection for *technology*, the *precise* truth is of less consequence, see section 1.3.2. The practical parallel for our case of technological introspection is the way that we can interpolate over the voids in the introspective report, see sections 6.3.4 and 6.3.5.

Witness accounts vary by the societal context. But even errors (or forgeries), once detected, can tell us much about the society that produced them (*Ibid.* p. 105). Ideas that prevail in a society, like the belief in German cunning in WW I France, led to an overestimation of their intelligence capabilities. Likewise, if you look at the naïve introspection of the pioneers of AI (see section 4.3.1) you will see that they attest that they think mathematically. This confabulation in itself is a testimony to the

mathematical prejudices of the scientific community in their times (a prejudice that remains with us today). Also in the case of the intuitions of cognitive psychology, testimony tells us often not what the witness saw, but what his society thinks it natural to see (*Ibid.* p. 107), see also (Nisbett & Wilson, 1977).

Summarising the accumulated wisdom of historiography as it applies to our case of introspection for technological human-like AI:

1. There is no mechanical logic of critical examination of testimony (*Ibid.* p. 110, 116). But we do not need a clear mechanical criterion for *truth* as such. We only need some social (perhaps economic) criteria for *utility*, and we have some – empirical testing and market economics - see section 6.5.

2. About the temptation to throw away all of subjectivity as non-scientific (section 4.2), we can adopt Bloch's advice that '*It is always disagreeable to say "I do not know. I cannot know." It must not be said except after an energetic, even a desperate search*' (*Ibid.* p. 59-60). But once we despair, we have a "get out of jail" card – since we are dealing in a technology that aims to *approximate*, we can interpolate the voids in our knowledge (see section 6.3.5).

6.3.2 Looking / listening for

In acting in different capacities, or in different roles (see section 5.3.1 about roles), we also *look for or listen for* different things (Cole, 1973; Winograd & Flores, 1986, pp. 9, 50, 57, 63). An example of the different types of "looking for" could be looking at the floor for a dropped coin, vs examining the floor for structural damage. The externally observable actions of surveying the surface can be identical, but the intent, the mental activity, and attention are entirely different. This is true also of our "internal looking/listening" in introspection. As Schwitzgebel (2012, sec. 1.2) notes, we can introspect either attitudes such as beliefs and desires, or conscious experiences such as emotions, images etc. The selection of what we listen for is volitional, we can choose to listen, together with the classic-AI people (especially the knowledge engineers) for beliefs and the application of rules, like in the case of the General Problem Solver (GPS) (Newell & Simon, 1961b). We can also focus on lower mechanisms.

Any report (not only introspective reports) beyond describing the subject-matter also

addresses a specific **audience**. On a trivial level, a bilingual person will issue his reports in one language at a time, the language of his audience. But also, any report will be couched in terminology that is expected to make sense to the audience: Stream-of-consciousness, data-and-algorithms, beliefs-and-doubts, whatever the reporter expects the audience to expect. Here is one of the places where Simon and Dreyfus completely talk past each other (see sections 2.1, 2.2): Simon is an objective scientist in the tradition of Watson, while Dreyfus is a phenomenologist, committed to subjectivity. They both listen for different things in their research, and address different communities (see chapter 2, and section 3.3.1.5).

So when we watch our own mental processes, or listen to ourselves thinking, we are always looking for something – there is no neutral observation. We also look to issue some report (if only for ourselves). What should we look for? In what terms should a report be made? We are looking for AI designs. A design (in this context) is a formalisation of a **process** of data processing, so we are looking for information processing processes in our mind. *That* is what we are looking for. But there are many pitfalls in the way: our favourite theory (e.g. that humans are rational) can interpose itself into our observations. It is important to "**bracket**" our beliefs and try to see the mental processes as they are (Gallagher & Zahavi, 2012). One of the most important things is "keeping our eye on the ball" of human thought as we truly experience it (insofar as possible) rather than how our culture might think our thoughts *should be*. That is how this approach differs from classic AI. However, we also need to "keep another eye" on programmability – or we will fall into the same trap as Dreyfus and his phenomenology, producing wonderful reports of no technological application.

A possible question arises of how can the enculturated mind look for or at the un-enculturated mind? There are two answers here:

1. The first answer is that there is a category error in this question, as if the "enculturated" and the "un-enculturated" minds are two different entities "inside our head", and there is a problem for one to access the other. The reality is that being enculturated or not is a matter of degree, or a matter of level. In a sense all we have is the un-enculturated level – and it goes to great lengths to "behave itself", i.e. to produce behaviour that is socially acceptable, enculturated. So it is

one and the same mind, just two levels – there is no access problem.

2. The second answer is that we should observe the un-enculturated mind "carefully". We should use all our skill, refinement, patience, openness etc. (all enculturated properties we hopefully have) in order to listen to the un-enculturated mind as truthfully as possible, and use our (hopefully) good command of language and writing to produce as accurate and honest a report as we can.

6.3.3 Pollution

So our listening determines what *kind* of introspection we do (beliefs, processes, sensations). It also colours the content of what we detect in introspection. Listening while being committed to some prior theory or model can lead to pollution of our introspection by our prior commitments.

An extreme example of what I call "polluting" introspection with prior theoretical commitments (in the context of an experiment) is beautifully recorded in (Newell & Simon, 1961a). This paper presents an experiment where the authors asked a student to solve a formal problem (while "thinking aloud", see section 3.3.3.1), given certain symbols and manipulation rules. They "*asked the subject to talk aloud about what he was doing-'what he was thinking about.'* " (*Ibid.* p. 2012), and they recorded the entire session. Only a few lines into the (quoted) session protocol the experimenters ask the subject "*Applying what rule?*", rather than a more neutral instruction such as "What are you thinking now?". The assumption that the subject is thinking in terms of *rules* is polluting the evidence, even when the explicit instructions were to report *all* their thinking.

However, the pollution can be more subtle, and self-induced. I would suspect that anyone saying something like "*For me mathematics has always been the language of thought. I don't know precisely what I mean by that... Mathematics – this sort of non-verbal thinking – is my language of discovery*" (Simon, 1996a, p. 106), is not being honest with himself or with us as his audience. Was mathematics *always* his "language of thought"? Even when he was 5 years old? Again, in introspecting for anthropic AI we need to get behind our "western, modern, well-trained and adult" thinking, and aim for the underlying mechanisms.

How can we defend against such pollution? First, we can make an effort to not project our theories or other cultural artefacts onto our introspection. This can be done, *inter alia*, by noticing our theories "coming to mind" and either setting them aside calmly, as a quiet choice, or turning our attention to how our thought processes specifically *deviate* from what such a theory might expect.

Next we can critique our introspections after the fact: Would such a mechanism produce the results we observe? As an example if "mathematics" as such were indeed someone's entire "language of thought", it would not produce mistakes, but we all do make mistakes. But ultimately we need no *steadfast guarantee* against such pollution. Different introspections, good and bad, pristine and polluted, will simply produce a larger variety of designs. These designs make no truth-claims, but are only candidates for a technology. As explained in section 4.7 point 7, let a thousand flowers bloom.

For anyone who still feels that the quality should be improved, there are detailed guides from the era before introspection fell out of favour, see (Schwitzgebel, 2004), and there is a whole field of modern research (Froese, 2011; Jack & Roepstorff, 2003). Again, these may be useful, but there is no necessity in taking them too seriously. Mistakes are acceptable – luckily for us, we are operating in the context of discovery, *for technology*.

For an explanation why introspection would be a good source of ideas for anthropic AI designs, see chapter 5. *Pristine* (non-polluted) introspection is better because the pollutions come from cultural artefacts, and we are aiming specifically for anthropic AI (see section 3.2).

Again, somewhat polluted evidence is not catastrophic, since we are exploring the space of possible AI designs. But in aiming for novel anthropic designs, we should at least try to avoid such unnecessary interference from over-optimistic over-mechanistic theories. A useful distinction in avoiding the tried, tested and tired theories of yesteryear is the distinction between knowing how and knowing that (see section 3.2.7). The very idea that "knowing that" is a fundamental part of thinking is a western, adult notion.

6.3.4 Introspection: is it above or below the culture line?

There is another problem, or seeming contradiction: When we introspect, often we come up with cultural products, like "I use mathematics" (to paraphrase Simon and others).

But aiming at building anthropic AI, we should look for the mechanisms that are *below* the "culture line" - for the basic *human* abilities and skills, that *underpin* culture. See section 3.2.5.1, where I cast doubt on the very concept of "layers" or "levels" in the mind.

Introspection, as Nisbett & Wilson (1977) argue, shows us the *products* of some unconscious, innate processes, so in a sense we can see *what* we are thinking, but not *how*. Here again I must protest that in terms of technology this is not a problem: in seeing "what" we think, with the ever-better resolution of practised introspection, aiming always to see more (see section 6.3.6), we will have enough of a picture of *what* is going on to interpolate (see section 6.3.5).

So are we introspecting at the cultural or subcultural level? Insofar as possible, we aim to introspect at that "boundary", but again insisting too much on exactness here is a fool's errand: it is not the distinctions we are ultimately after, it is the technologies. More discussion of this point is in section 5.2.5.

6.3.5 Interpolation and approximation

6.3.5.1 The holes in introspection

The consensus in cognitive science seems to be that introspection may tell us *what* is being thought about, but not *how* this thinking is accomplished (Nisbett & Wilson, 1977).

This is seen as a problem for Introspection, but that problem is in using introspection for the science of psychology, rather than for technology. If we know only the outline of *what* is accomplished without the *how*, we can substitute whatever technical trick we have (in our skills as programmers) to achieve the same in a computer. For example, we do not fully understand how long-term memory works in humans, and the introspection that "I just recalled my first day in school, the weather was dreadful!" - does not help us explain *how* the memories are stored in the brain. But in technology we can be far more relaxed – if we need some long-term store of information, we can use an SQL database. Here we see the "cash value" of the insistence that technology and science have different criteria for truth, see sections 1.3.2, 1.4. We need not obsess about the true mechanism of memory with the cognitive psychologists – we can just go ahead and write code. The gaps between the different moments in introspection ("trying to recall…

recalled!") are a problem for science and not for technology, for psychology and not for AI.

My proposal of using introspection in AI design is not a proposal for a new "do-all" technique like expert systems or "deep learning" with which often are deployed as an entire solution. My proposal is to use introspection to design systems that use, and rest on top of any and all previous technologies. This is very much like Minsky's "scruffy" AI (Minsky, 1991). So we should use introspection for the overall design, and perhaps for some of the components – but as technologists we should not shy away from using existing techniques as part of the design.

Note (following section 5.2) that also when humans teach each other skills (like making a cappuccino with an espresso machine) the teacher does not teach how to move one's hand, or how to lift the milk-canister. The assumption is always that more basic skills that can be used *pre-exist*. In AI some of the more basic skills may be implemented using an introspection-based algorithm, some can be implemented using some other AI, and some can just be hard-coded.

6.3.5.2 Opportunistic approximation

When we need to implement some mechanism gleaned from introspection, we often do not have enough information on what the mechanism does *precisely*. For example, we may forget something or overlook the best option in some fraction of the times we try to achieve a task. We can use crude approximation, like "50%", and later tune that parameter if the result is not a good match of the observed introspection or behaviour. Moreover, we can sometimes match a phenomenal process that needs to happen "occasionally" with some computational process that would be expensive (say in CPU time). An example of that is given in section 7.5.3.5. These are just conveniences, and as long as the AI works and produces credible behaviour, they are OK. Again, we are not doing science.

6.3.5.3 Analogue cannot arise out of digital

An objection may arise that the subtlety and fluidity of our subjective mind cannot be captured by the 0s and 1s of a computer (similar to Dreyfus's objections as summed up in section 2.3.2). Though it may be true that humans are essentially analogue and

computers are essentially digital[19] and therefore cannot be the same, we still can approximate analogue phenomena to an arbitrary precision, especially with the current availability of virtually unlimited computation power “in the cloud”. Just as we can implement floating-point numbers as substitutes to real numbers, and we can add precision by adding bits, and just like we can simulate the earth's atmosphere for weather forecasts by simulating the physical conditions in “air cells”, so we can make an approximation of the fluidities of human subjective experience.

6.3.5.4 Being analogue does not mean it is not digital

On the contrary, one could speculate that maybe humans' non-ideal, **informal behaviour is nonetheless produced by an underlying ideal mechanism**, perhaps in a similar way that deterministic behaviour by a computer can be used to simulate and predict seemingly chaotic systems, such as the weather. My response is as follows:

1. It is highly unlikely that there is such perfect order as Case Based Reasoning (see section 7.2) or Intel processors underneath our rather non-formal experience of ourselves dealing with the world. There seems to be nothing in the brain that operates digitally, or at a sufficient frequency to “simulate” our informal experiences (see section 2.3.2).
2. If there were even a likelihood of such an underlying order, then the onus to show that such an order exists would surely be of those who propose its existence and not on those who deny it (by Occam's razor, or in analogy to Russell’s teapot (B. Russell, 1952)).
3. Regardless of whether such an order ultimately exists underneath the seeming informality, such a mechanism for producing chaos out of order it is not visible to us in any form that can be used to base technology on it.

6.3.6 Multiple iterations, multiple mechanisms

In using introspection as a basis for AI, we may introspect even briefly, not too thoroughly, stop and implement the model we came up with, and then come back and refine our introspection and our model again, and then refine the code. There is no necessity to produce a complete tome of phenomenology (or several) before we start

19 Brian Cantwell Smith would protest that computers are in the world, subject to the same physics, and are only “ideally” digital. Though interesting, this point does not affect the argument here.

coding. This is a point of difference from Dreyfus (2007).

As we will see in section 7.1, (one could argue that) fuzzy logic was based on introspection, saying that the boundaries of concepts are not clear-cut but fuzzy. Zadeh did not deepen this introspection, nor did he broaden it. Deepening it would mean further exploring the ways that concepts behave in our subjective experience, and broadening it would have to do with adjacent mechanisms, say memory, or action choice.

Zadeh used introspection for this one element, and fell back on to logic, mathematics and the pre-established tradition of expert systems. I would argue that if we want to create anthropic AI, we need to introspect multiple mechanisms, and *not* include any artefacts of a specific culture, like mathematics or logic. We most probably need multiple novel elements, conjoined in a way that respects our introspective observations, not some “neat” architecture (Minsky, 1991). These elements can be added gradually, interspersed with experimentation for feedback on how each step of the development of the AI design works, and allowing further introspection at each stage, looking at how our own thought-processes differ from the model.

6.3.7 Personnel

In a sense, this thesis flies in the face of the traditional division of skills and mindsets between the hard sciences (STEM) and the humanities, and also is distinctly non-cooperative with psychology in its quest to become an exact science. But this is not just a theoretical point: in terms of personnel, if one wants to develop anthropic AI using introspection, perhaps STEM education and programming skills are not the principal skills that are needed. If indeed introspection is key, there is a need for people who are good at that. I would wager that people with a sense of poetry, drama, literature etc. may be useful members of a team developing AI. Such a project needs people who are more at home with the soliloquy than with the compiler. In a sense this is a direct reply to Snow (1964). We have done mathematics and cognitive-theory based AI for long enough. It is time to try something radically different, rather than “returning to cognitive science” as Langley (2006) had it.

That may have been a little overstated. In any team where software is developed, programmers are key. But the “architects” of the software must be informed by

introspection rather than by the latest software development fad. It would of course be a good idea to have the entire team be composed of people who are good introspectors, knowledgeable about all exiting AI techniques, and also good programmers. It is unlikely that any team will manage to recruit more than one or two of these fully interdisciplinary workers, if that. Instead, it is enough that the introspectors have a vague idea of programming, so that they produce models that can be at least approximated into some software design (see section 6.3.5). The software architects, doing the design, should at least have a healthy respect for the introspection process, but need to have a fully professional grasp of programming, so that the programmers would fully understand what is required software-wise. Again (as we saw in section 3.3.1.4), "doing" subjectivity or introspection is not one thing, a box to be ticked, but is an ongoing and iterative process. This should not be taken lightly.

Interdisciplinary thinking is not just desirable to "compete with the Russians" as Snow (1964) demanded. Interdisciplinary work is a positive requirement for developing anthropic AI. These different disciplines may be found in the same person, or in a team that works well together.

6.4 Project expectations

Consider the (currently impractical) idea of building AI by simulating every cell and interconnection in an entire human brain (As in the blue-brain project (Markram, 2006)). Since the AI would be constructed according to some scientific model of the brain, that would initially be quite inaccurate, we should not expect the mind emerging out of such a simulated "brain" to necessarily be sane, of sound intelligence, or interested in communicating with us humans. This arises from many possible causes: many parameters will be inaccurate, the simulated brain would hardly undergo a normal social development, etc. (Idan Segev, personal communication, 2011). In such a set-up, we would be delighted if we get a mind capable of *any* learning in even very few of the cases.

Similarly, in an AI system that is sufficiently "low level" even if not as low as the cell level, we should expect a relatively *low success rate* in engaging with the environment in a way that would be meaningful *to us*. In aiming for anthropic, i.e. sub-cultural AI, we are in the danger zone, courting those difficulties. Perhaps we will need two distinct

phases of development – the one where the anthropic model is being developed "for its own sake", brought to a functional level "in the lab", and a later "implementation" phase where only the better specimens of the original model are actually used in implementing practical technology. Only in this implementation phase would a normal technical evaluation make sense.

6.5 Testing and evaluation

In section 1.2.1 I mentioned that this thesis is in a sense the middle volume in a 3-volume project, with the third volume being a proper technological exploration and evaluation of algorithms. This section gives a sketch of the evaluation methodologies that should be used, while chapter 7 gives a taste of the possible technologies.

Recall that the purpose of using introspection is both to generate a more anthropic AI design, and to broaden the bases for AI development. As a technology, Anthropic AI must ultimately pass muster as being fit for purpose, workable, and marketable (see section 1.3.2). However, in the development stage, if one is interested in how anthropic their technology is, the following points may be of use:

The purpose of anthropic, non acculturated AI is to create human-like systems that can learn as flexibly as humans are flexible, without being pre-committed to a specific way of doing things. So we are interested in how human-like an AI system is, compared to other systems.

A conservative evaluation of any design should be empirical. As alluded to in section 6.2, any evaluation can follow an objective or a subjective (quantitative), path. Further (and less conservatively) some qualitative feedback may be of use:

- In the objective path, both humans and introspective-based systems are put in similar circumstances, and measurements are taken of various parameters. The more **similar** the system is to the human the better it models a human, This is similar to cognitive simulation (Sun, 2008) in method though not in intent – the intent of cognitive simulation is to produce better scientific models for psychology, while in our case the intent is to produce technology.
- The subjective path for evaluating an introspective system would involve producing (say) a video of the performance of various algorithms (and perhaps

also of humans tackling a task) and asking a sample of disinterested observers to give their impression of “how human or machine-like” each video seems. The data collected would be processed using standard interview-data methodologies as in the social sciences.

- Qualitative feedback could also be of interest, collecting the comments of the observers as feedback for the developers.

The choice of evaluation methodology would be influenced by the aims and circumstances of the project.

7 Examples of introspection being used for AI design

Table of Contents

7 Examples of introspection being used for AI design....161
7.1 Fuzzy logic....164
7.2 Case based reasoning (CBR)....166
7.3 AIF0....168
7.3.1 Introspection....168
7.3.2 Implementation....169
7.3.3 Example run, statistics....170
7.3.4 Discussion....173
7.3.4.1 Details and parameters....174
7.3.4.2 Why this is more anthropic....175
7.3.4.3 Similarity....175
7.4 AIF1....176
7.5 AIF2....177
7.5.1 Introspection....177
7.5.2 Introspective model:....178
7.5.3 Software design....179
7.5.3.1 Sequences in software....179
7.5.3.2 A novel data type....181
7.5.3.3 Decision process....183
7.5.3.4 More details of AIF2's implementation....184
7.5.3.5 Dynamics of the scenario table....185
7.5.3.6 Initial conditions and decisions....186
7.5.3.7 Further Parameters....186
7.5.4 AIF2 Example runs....187
7.5.4.1 Learn 1....188
7.5.4.2 Learn 2....189
7.5.4.3 Learn 3....189
7.5.5 Discussion of AIF2....189
7.6 Consequences of the examples....190
7.6.1 AIF is more like CBR then like reinforcement learning....190
7.6.2 The "sequence" data type....191
7.6.3 Dynamic symbols....191
7.6.4 How AIF2 is Gadamerian....192
7.7 Examples of introspection being used for AI design: summary....193

This chapter presents five examples of AI: the first two, fuzzy logic and CBR, are

existing technologies which will be used for illustrating some preliminary points, while the next final three examples are based on the proposed methodology for developing anthropic (human-like) AI. The methodology presented in the advanced examples is in a sense a compromise between Classic AI and Phenomenology. Like phenomenology it follows the subjective, introspective angle. Like classic AI, observations are (perhaps grossly) simplified in order to make them programmable, but not so much as to make the resultant mechanisms phenomenologically non-credible.

The first example, Fuzzy Logic, will demonstrate a minimal case of using introspection for AI. The second, Case Based Reasoning, will show a more developed example of how this design could have evolved using introspection (thought historically it is not clear whether introspection was involved). The third example will be "AIF0", my first experiment in developing introspection-based AI, followed by AIF1 (a failed effort) and AIF2 – a successful and interesting design.

The concern being addressed is technological, not scientific, so the very question "is it correct?" is less to the point than "does it work?" (see section 1.4.4). This calls for a few clarifications:

1. The introspection data presented in the examples is derived just from ***my*** **introspection**. I present it "as is" without any attempt to argue for it, amongst other reasons because it is not clear what sort of argument *can* be presented for phenomenological / introspective data. Regardless of the accuracy or veracity of the introspective reports below, what I am arguing for is the *methodology* that uses introspection as a basis for AI designs. Anyone can use this methodology to design AI based on their own introspections, let a thousand flowers bloom (see point 7 in section 4.7).

2. Building on that point, since the examples presented are only *examples*, from one person's introspection, no claim is being made that the resulting designs are **good** in and of themselves, and therefore I will not present comparative data trying to prove that any of these example designs are better that any existing design by some objective criteria. Any competitive evaluation would fall outside the scope of this volume (see section 1.2.1). The claim is only that it is ***plausible*** that they would be more human-like, by using an anthropic (section 3.2),

introspection-based design process (chapter 6). The focus of this thesis is showing a new *methodology,* a different way of *developing* anthropic AI (but see section 3.2.6) based on the deliberate use of introspection (this is also a compromise between classic AI and phenomenology). I am not arguing here that the example designs are fit for any specific purpose, only that they are *likely* to be more anthropic (See sections 3.3.1.1 and 5.2), and demonstrate the fecundity of the approach (see section 8.2).

3. The focus in AI design is **technological**, and the criteria of success for technology are lower than for scientific truth (see section 1.3.2). However, once a design exists, regardless of its source (see "context-of-discovery", section 4.2.4), a model based on such a design could be proposed as a scientific model (theory) in psychology, similar to how some simple neural-networks are proposed as theories in cognitive psychology (e.g. (Altmann & Dienes, 1999)). People who want to resurrect subjective psychology from J. B. Watson's (1913) blows may hold hope for such theories. This should be contrasted with Wheeler's (2005) approach, which sees AI as "the intellectual heart of cognitive science" - he is interested in the science first. Here it is technology first. That changes the level of truth we need to ascribe to our models, and allows much more freedom. We are not doing science here, in any case not directly (see section 8.3.1).

The first two examples, fuzzy logic and case based reasoning (CBR), are given as a "warm up", to illustrate some points. The historical evidence as to the degree to which introspection was used to develop these designs is very partial and at times contradictory. Rather than worry about these historical points, I will present them *as if* there were a deliberate use of introspection in both cases, in order to use them as examples.

Environment

In all the following examples other than the first (fuzzy logic) the environment is a game-like situation, where the AI can take actions in the environment (from a predefined set) and gets feedback in the form of a score.

Actions come from the range of available actions that can be performed by the machine.

The outcomes of the machine's operation are given as **score** inputs. The scores allow measurement of progress towards one or more goals, and/or adherence to one or more principles. Scores may be provided from the environment, or from a subjective source such as a human observer.

7.1 Fuzzy logic

Note that the first two examples are existing AI designs, that are here for illustration purposes only. The points being illustrated are three: 1. how introspection could be used, in a minimal way, and 2. How we can step away from the boolean, overly rational AI, and 3. mentioning some mechanisms that are "nearest neighbours" of the original designs to be presented later.

Note that the historical facts about whether Fuzzy logic (here) and CBR (in section 7.2) were indeed derived from introspection are unclear – in both cases there is evidence both ways. The conclusive historical facts are beside the point here - I present these examples *as if* they were derived from introspection, as illustrations.

-

Fuzzy Logic is the first example because it shows, in one isolated and clear case, how introspective AI could work, and how it did, in a *minimal* way. The fuzzy notion arose from Lotfi Zadeh's (1965) self-observation that in human concepts, the boundary between membership and non-membership of a class (or set, or concept) is not a square-wave or all-or-nothing affair.

The "fuzzy" notion is the idea that things need not be 100% members or non-members of a set, or category, or concept. The example in Illustration 7.1 shows that a specific temperature, say 10˚C (shown here as the vertical line) can be seen as being 90% cold, and 10% warm. So the temperature "10˚C" is only a part-member in the concept or set "cold", and likewise for the concept or set "warm". One of the applications of fuzzy concepts is in formulating rules for expert systems, allowing words like "a little" "somewhat" "very"

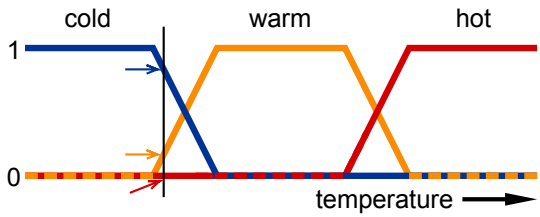


Illustration 7.1: Fuzzy Logic *(Source: Wikimedia)*

etc. to be given numerical meanings, and expert-system rules to say things like “when the boiler is somewhat warm, do X”. This is language closer to the (human) expert's language. The eventual decision processes (de-fuzzyfying) vary between implementations (McNeill & Freiberger, 1994). To clarify the terminology, fuzzy *sets* allow parsing of notions such as “somewhat warm”, while fuzzy *logic* is a design for combing such notions, similar to standard logic's AND, OR, etc.

To see how Fuzzy Logic can be seen as a minimal case of “introspection-based AI”, let’s trace the process of turning an introspection into an AI design using the terminology of Section 6.2, explicitly referred to here in **bold type**. Zadeh (supposedly) aimed to represent the way he *actually* thinks (phenomenologically) rather than the way that a logical, mathematical or scientific mindset would urge him to think (this is the intention to introspect). Out of his **introspective vision** Zadeh chose to concentrate on (or **listen for)** the boundaries of concepts, or “sets” as he called them in his paper (1965). Regardless of any complexities in his introspective vision, he chose to **articulate** his **model** of his observation as allowing each object to be a member of any concept or set to an *extent*. This extent is expressible as a percentage or as a real number between 0 (non-member) and 1 (full member).

Later (in the **software design** phase) Zadeh chose to approximate the boundaries of the concepts using simple mathematical curves, such as the straight linear function seen above in Illustration 7.1. The compromise involved with this approximation for the sake of **programming** is clear – the diagonal lines in the diagram above are unabashedly mathematical (this is a case of opportunistic approximation, see section 6.3.5.2) and therefore not precisely human or introspective, but are still significantly more representative of the human situation than assuming an all-on-nothing (square wave) membership of a category.

Again, we should not err (with Dreyfus) into an excessive belief in phenomenology or the phenomenological literature: the idea is to introspect in order to create usable designs in a finite (development) time. It is fine to approximate. On the other hand, Zadeh is an excellent example of how one point of introspection is taken, and then embedded in a mathematical framework (set theory, logic…). Zadeh is not doing full-blooded introspection – he is committing one little introspective sin, so to speak, and

hurries back to the safe shores of mathematical orthodoxy. So yes, he used introspection as basis for AI, but did so sparingly and shyly, see section 4.3.

Some may object to any discussion of fuzzy logic, in that it has been shown to be a specific case of statistical AI, and therefore of no lasting contribution. That would be missing the point here – here we are discussing the method of invention. The fuzzy notion, like most other notions, is limited, perhaps already obsolete. It is here to demonstrate a minimal usage of introspection, and how it was done shyly.

7.2 Case based reasoning (CBR)

CBR has its roots in scripts, dynamic memory, and other cognitive theories (Nilsson, 2010, pp. 400–402; R. C. Schank & Abelson, 1977; Roger C Schank, 1982; I. Watson, 1999) one could tell a (perhaps reconstructed, perhaps speculative) story about how CBR was arrived at, in order to illustrate the point of introspection-for-AI. As we saw in section 4.3.1 (in an interview with Turkle) Roger Schank explicitly admitted to using introspection – however he expunged any mention of introspection from his published AI papers.

CBR is of interest because it exemplifies how some of AI *was* based on introspection (very poorly, details of what is better introspection are in section 3.3, chapter 6), and because it is a stepping stone towards presenting the next, more interesting (and original) examples.

CBR (arguably a methodology rather than a specific algorithm (I. Watson, 1999)[20]) attempts to solve every problem (or situation) the agent may encounter by looking into a database of previously encountered problems and solutions, and selecting the best solution available. In some cases, the solution is adapted for the current case, before it is executed. The slogan for CBR is the 4 “Re”s: “retrieve, reuse, revise, retain”. This design has had some success, and has many variants (I. Watson, 1999).

A reconstructed/speculative story on the origin of CBR could go like this: while playing

20 I. Watson argues that CBR is a methodology rather than an algorithm, since in its purest form it advocated only the 4 “Re”s, retrieve, reuse, revise, retain. The details on how to do each of these steps remained unspecified. However, if one wanted to be pedantic, any algorithm that includes even something as simple as “add 1 to a” is underspecified, in that the exact behaviour of a in terms of overflow is left unspecified. In our case here this point is not so important. I use the term “methodology” in this thesis for an approach for developing *new* AI designs, and the term “design” for algorithms and families thereof.

the **role** of competent administrators (in the spirit of (Simon, 1976), see section 5.3.1), the inventors asked themselves “what would *I* do?”, “How do *I* solve problems?”. They came up with the answer of re-using solutions that have been accumulated in their own memory from previous encounters. The design does not specify where the original database comes from, or how new solutions are generated for problems never encountered before – again like a well trained administrator, the assumption is that the computer has already seen all the relevant cases, and can make minor adaptations. Hence, at its most primitive, CBR is little more than a directory of solutions.

Note this latest term, “**solutions**”, which suggests that the following assumptions are being made:

1. That the world in which the agent is operating is made of distinct “problems”, presented one by one;
2. These problems admit of solutions, that these solutions are clearly and obviously distinct from non-solutions, which would be labelled *wrong*, and as useless behaviour;
3. That the database already has such a solution in store for each problem to be encountered, or a solution that can be readily adapted for any presented problem (the “revise” stage).

So we see that CBR is made in the image of an idealised, rational, administrator. Solutions are distinct from non-solutions, there is no matter of degree of “goodness”; where all the assumptions of this design are correct in the world in which the AI operates then the design would produce ideal behaviour – bounded rationality (Simon, 1955). There is nothing wrong with all that, except it does not meet *our* agenda of producing anthropic AI. See also section 2.3.3 (point 3) for Dreyfus's distinction between “situation” and “problem”.

The possible objection that non-ideal, perhaps anthropic, systems may be underpinned and implemented by ideal systems is handled in section 6.3.5.4.

The assumption that solutions are clearly and immediately separable from non-solutions is unrealistic, and will further fade away as we progress to a more introspective designs, such as the next example.

7.3 AIF0

This example demonstrates moving away from the mathematical towards the more introspective. It is a first step into technical novelty. There are two more to come. As a point of nomenclature, I will use “select” for picking a *subset* out of a bigger set (a bit like SQL uses that word) and will use “choose” for the *one* final decision of an algorithm in a specific round.

Let's refine CBR. I will describe the process of introspection (the contents of introspections are underlined), and the processes of software design, running and results, and will later discuss the implications of this example.

7.3.1 Introspection

One could **observe** that in everyday life, I often use suboptimal solutions for problems, not because I have not encountered a better solution, and not because it is “good enough” (Simon's satisficing) but just because it is not in my nature to always do the perfectly correct thing – either through ignorance, confusion (mis-execution), or through playfulness or exploration (remember we are *not* aiming for emulating the well-trained scientist or soldier). So in terms of introspection, I do things that I know (or hope, or believe) will work best, but I do not necessarily choose the best of the options I know of. Sometimes I just guess, or do something new, even if I know a good response to a situation. I also sometimes mis-execute my intentions.

Here is an **articulation** of the following innovations over CBR, a **model**:

1. Not one “solution” is selected, but several, and
2. One of these is chosen with some random element.
3. The very notion of “solution” (as in a 1-0 solution-on-not sense) is gone, and is replaced by the notion of “the best few we know”. Therefore the assumption that we have a correct solution for every prospective problem, or even one “good enough” solution is relaxed and replaced with “the best few we have”.
4. Unlike CBR where “the” solution (or “the best” solution) is chosen and used, in AIF0 there are steps. First multiple similar cases are selected (perhaps up to some threshold of similarity), and next these similar cases are sorted by expected outcome, derived from the episodes’ score as found in the memory bank. In a

relatively rare case, a random action is performed.

5. A score, or feedback, is collected from the environment.

7.3.2 Implementation

In the process of **software design** we must approximate the introspective model using a software mechanism, so in this case we can use the following approximation for fallibility and playfulness of the human: use the best “case” solution only half of the time, and the second best “case” in another ¼ of the cases, and so on for the next 1/8 and 1/16 of the cases. In the last 1/16 of the cases we can have the design choose an output at random from the repertoire of possible responses (this is an example of “interpolating” a mechanism, see sections 6.3.5.1, 6.3.5.2). This mechanism facilitates learning, so this design can bootstrap its own knowledge bank, without any prior knowledge.

Pseudo Code:

- For every situation:
    - recall all similar situations from the past, (similarity can be crudely defined as equality for the time being)
    - Of these similar solutions, select the top best (outcome) 4 cases that were used, and sort them by score,
    - i ← first case,
    - while i is a valid case:
        - Flip a coin (50% chance).
        - If heads,
            - choose the i'th case (goto “DONE”)
        - else
            - i ← next case,
    - When the “while loop” ends, choose a random action
    - DONE:

- Perform the chosen action in the “world” (which is probably a micro-world, simulated), collect a score, store the case, and repeat from the start for next input

All that is left of the original introspection (that was not in CBR) is

1. that we select *multiple* cases, and
2. choose from these *non-deterministically*, and
3. allow for some ongoing degree of random choice.

This is a stepping-stone, and this is a technological project, so adding three features at a time is a reasonable step.

Note also that the whole issue of “**similarity**” is being brushed neatly into a subroutine (as it is in CBR). The following example will use identity as a crude form of similarity.

7.3.3 Example run, statistics

This algorithm was run in a world consisting of a game where each input, A, B, C, or D should be matched with 1, 2, 3, or 4 as output, respectively. A successful match scored 1, an unsuccessful match scored 0.

Each “case” starts with the algorithm getting an input of either A, B, C, or D. The algorithm produces an answer, 1, 2, 3 or 4. Below is a trace from an example run.

Explanation of the trace: Please follow as an example the highlighted line number 45. Every iteration number is followed by the input ('A' – 'D'), then the 4 “best matches” (selected options) are presented (represented by their iteration number) from which the algorithm will later choose. These four “best matches” are represented by the “iteration number”, so in our (highlighted) example the input “C” is similar to all previous instances of a “C” input, and these are sorted by score, and the best case is 18, followed by 19, 36 and 39 (in all of which the program scored 1). Note that in the first few lines of the trace, where there are no precedents to follow, an index of “-1” is shown. Next in the output line is the index of the chosen option, 0-3 (for the four possibilities) or R, if a random action was selected. Note that when there are few options available (near the top) then the random option is common. Next the output is shown and the score.

```
ITER  Inp.  Options to consider (4), selected      Score
   0, in B: ops(  -1,   -1,   -1,   -1)-> R out 4, S= 0
   1, in C: ops(  -1,   -1,   -1,   -1)-> R out 1, S= 0
   2, in B: ops(   0,   -1,   -1,   -1)-> R out 3, S= 0
   3, in A: ops(  -1,   -1,   -1,   -1)-> R out 2, S= 0
   4, in D: ops(  -1,   -1,   -1,   -1)-> R out 3, S= 0
   5, in A: ops(   3,   -1,   -1,   -1)-> R out 2, S= 0
   6, in C: ops(   1,   -1,   -1,   -1)-> R out 4, S= 0
   7, in B: ops(   0,    2,   -1,   -1)-> 1 out 3, S= 0
   8, in C: ops(   1,    6,   -1,   -1)-> 1 out 4, S= 0
   9, in D: ops(   4,   -1,   -1,   -1)-> R out 2, S= 0
  10, in A: ops(   3,    5,   -1,   -1)-> 0 out 2, S= 0
  11, in D: ops(   4,    9,   -1,   -1)-> R out 1, S= 0
  12, in A: ops(   3,    5,   10,   -1)-> R out 2, S= 0
  13, in B: ops(   0,    2,    7,   -1)-> 0 out 4, S= 0
  14, in B: ops(   0,    2,    7,   13)-> 0 out 4, S= 0
  15, in A: ops(   3,    5,   10,   12)-> 0 out 2, S= 0
  16, in B: ops(   0,    2,    7,   13)-> 1 out 3, S= 0
  17, in B: ops(   0,    2,    7,   13)-> 2 out 3, S= 0
  18, in C: ops(   1,    6,    8,   -1)-> R out 3, S= 1
  19, in C: ops(  18,    1,    6,    8)-> 0 out 3, S= 1
  20, in D: ops(   4,    9,   11,   -1)-> 2 out 1, S= 0
  ...
  35, in B: ops(   0,    2,    7,   13)-> 0 out 4, S= 0
  36, in C: ops(  18,   19,    1,    6)-> 0 out 3, S= 1
  37, in B: ops(   0,    2,    7,   13)-> 1 out 3, S= 0
  38, in A: ops(   3,    5,   10,   12)-> 1 out 2, S= 0
  39, in C: ops(  18,   19,   36,    1)-> 0 out 3, S= 1
  40, in D: ops(   4,    9,   11,   20)-> 1 out 2, S= 0
  41, in A: ops(   3,    5,   10,   12)-> 0 out 2, S= 0
  42, in C: ops(  18,   19,   36,   39)-> 1 out 3, S= 1
  43, in B: ops(   0,    2,    7,   13)-> 0 out 4, S= 0
  44, in C: ops(  18,   19,   36,   39)-> 0 out 3, S= 1
  45, in C: ops(  18,   19,   36,   39)-> 3 out 3, S= 1
  46, in A: ops(   3,    5,   10,   12)-> 0 out 2, S= 0
  47, in D: ops(   4,    9,   11,   20)-> 0 out 3, S= 0
  48, in C: ops(  18,   19,   36,   39)-> 0 out 3, S= 1
  49, in D: ops(   4,    9,   11,   20)-> 2 out 1, S= 0
  50, in A: ops(   3,    5,   10,   12)-> 0 out 2, S= 0
  51, in C: ops(  18,   19,   36,   39)-> 1 out 3, S= 1
  52, in A: ops(   3,    5,   10,   12)-> 0 out 2, S= 0
  53, in C: ops(  18,   19,   36,   39)-> 0 out 3, S= 1
  54, in A: ops(   3,    5,   10,   12)-> 0 out 2, S= 0
  55, in C: ops(  18,   19,   36,   39)-> R out 4, S= 0
  56, in B: ops(   0,    2,    7,   13)-> 3 out 4, S= 0
  57, in C: ops(  18,   19,   36,   39)-> 0 out 3, S= 1
  58, in A: ops(   3,    5,   10,   12)-> 0 out 2, S= 0
  59, in C: ops(  18,   19,   36,   39)-> 1 out 3, S= 1
  60, in C: ops(  18,   19,   36,   39)-> 0 out 3, S= 1
  61, in A: ops(   3,    5,   10,   12)-> 2 out 2, S= 0
  62, in A: ops(   3,    5,   10,   12)-> 0 out 2, S= 0
  63, in C: ops(  18,   19,   36,   39)-> 2 out 3, S= 1
  64, in B: ops(   0,    2,    7,   13)-> R out 2, S= 1
  65, in A: ops(   3,    5,   10,   12)-> 2 out 2, S= 0
  66, in A: ops(   3,    5,   10,   12)-> 2 out 2, S= 0
  67, in A: ops(   3,    5,   10,   12)-> 2 out 2, S= 0
  68, in C: ops(  18,   19,   36,   39)-> 1 out 3, S= 1
  69, in D: ops(   4,    9,   11,   20)-> 0 out 3, S= 0
  70, in C: ops(  18,   19,   36,   39)-> 0 out 3, S= 1
  71, in D: ops(   4,    9,   11,   20)-> 0 out 3, S= 0
  72, in D: ops(   4,    9,   11,   20)-> 1 out 2, S= 0
  73, in C: ops(  18,   19,   36,   39)-> 0 out 3, S= 1
  74, in C: ops(  18,   19,   36,   39)-> 1 out 3, S= 1
  75, in C: ops(  18,   19,   36,   39)-> 2 out 3, S= 1
  76, in B: ops(  64,    0,    2,    7)-> 0 out 2, S= 1
  77, in C: ops(  18,   19,   36,   39)-> 0 out 3, S= 1
  78, in B: ops(  64,   76,    0,    2)-> 1 out 2, S= 1
  79, in A: ops(   3,    5,   10,   12)-> 0 out 2, S= 0
```

Note that by the end of this trace (80 rounds) the machine has learnt about the mappings for B and C, but not for A or D. This will likely happen later in the run and the entire

mapping will be learnt. Illustration 7.2 shows an average of the scores of 10,000 runs of a 2,000-round game.

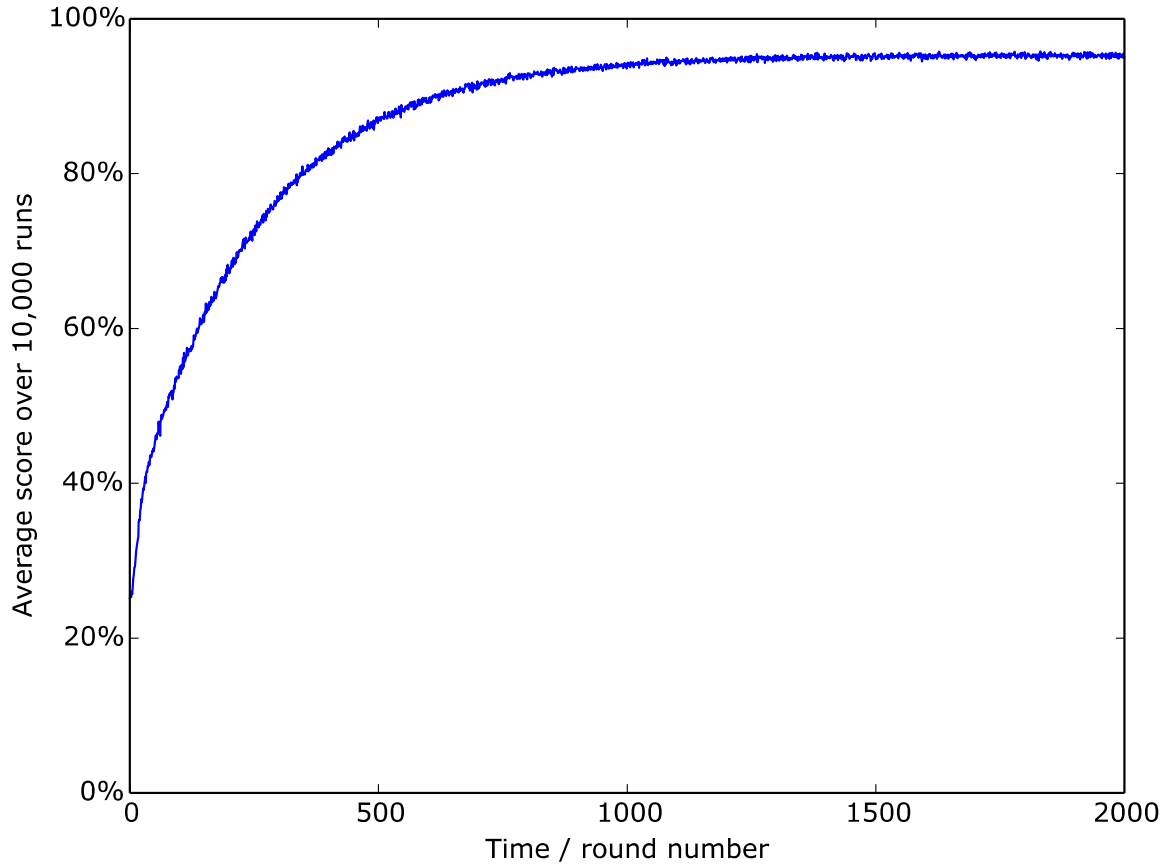


Illustration 7.2: Statistics on the Learning of AIF0 on the ABCD -> 1234 problem

Note that the theoretical maximum that this graph should approach can be calculated as follows:

Once the program has run for long enough to have 4 samples of a correct response for each possible input, it will usually choose the correct answer based on this accumulated experience. The chances of it choosing a random choice are 0.5^4 = 0.0625 = 6.25%. If the software chooses a random action, it still has a 25% chance of choosing the correct answer randomly, so the ultimate error rate should be 4.68%, and the success rate should be 95.32%, as we indeed see in the graph above.

Consider the following modifications:

- If the rules of the game (the correct mappings between ABCD and 1234) change midway through the game, this would lead to much confusion. A slight variation, making the "case number" part of the input that is compared by the similarity function, would allow the algorithm to recover from a change in the rules as it would prefer more recent "cases" since they would be more similar to the current case.
- Adding a random small "noise" to the score does not change the results noticeably, though it makes the "similarity" more like a real world scenario.

Recall, the purpose of this algorithm is only as a simple example to illustrate the wider point about the methodology. In subsequent section this will be built upon.

7.3.4 Discussion

Some of the observations in the introspection above could be reached using other, more objective, means. As long as an introspection gives us *some* new insight for a design that insight is worthy of consideration. In following introspection rather than mathematical correctness we move away from rational AI towards anthropic AI, away from classic AI towards subjectivly-informed AI, away from classic AI and towards Dreyfus.

In this specific algorithm we introduce *mistakes* (not choosing the best option, the 1/16 chance that the result will be random), but we gain the algorithm's ability to bootstrap its own knowledge, i.e. to learn, starting from an empty database. Moreover, we gain the ability to recover from a mid-way rule-change. Also, through introspection we uncover three distinct sources of sub-optimality, and these are reflected in the software design:

- **Ignorance** – we simply do not know a better response to a situation. This is reflected in the AI design in that the software can also be ignorant of the "better solution".
- Confusion (**mis-execution**) - "the spirit is willing, but the flesh is weak"[21] - like tripping – there was a full intention to do the right thing, and the execution failed. This is mainly reflected in the possibility that only the 2nd-to-4th options

21 Matthew 26:41

will be chosen. It is also reflected by the “random answer”, the fifth option in each decision.

- **Playfulness** - Humans (not well-trained soldiers) often do not go for 100% accurate performance, since they find it boring. Humans like experimenting – we could call it playfulness. Contrast a 2 year old with undergraduate students, and with elderly people: the older a person gets, the more they tend to refrain from experimentation and play-behaviour, and the more you can expect them to “behave reasonably”. This tendency to reduce exploratory moves (playfulness) with time can be reflected in the algorithm by gradually, during the run-time of the above design, moving the chance of using each selected case up from 50% to say 70%. The result would be that gradually during the game the chance of a random choice moved from 0.5^4 (6.25%) to 0.3^4 (=0.81%). This will allow for sufficient experimentation in the early stages, but will allow a better result later on – at the cost of the learning ability later on. This modification of this “**decisiveness**” parameter during the run is one of the variations possible with this design (see section 8.2). Playfulness is reflected mainly in the random action, and also somewhat in the possibility of choosing the 2nd-to-4th options.

7.3.4.1 Details and parameters

Not every aspect of the code reflects introspection directly, e.g. the notion of “flipping a coin” in this particular way is just an attempt to approximate the apparent randomness in the introspected process. Note also that the parameters “50%” (coin flip) for decisiveness or “4” (number of attempts) are **arbitrary**, and may be tuned to get different behaviours. Note also that these details are crude and mathematical, and have no basis in introspection whatsoever. This is part of the compromise between classic AI and phenomenology, and is in line with the precedent of fuzzy logic, where the fuzzy edges of the concepts are assumed to be linear, or described by some other simple function (see section 6.3.5.1).

A question might arise here as to what degree can these arbitrary **parameters** be tuned away from the initial values (50% and 4 here) before the AI method can no longer be considered “based on introspection”. Being interested in technology I view this question as scholastic (in the derogatory medieval sense) – as long as an activity (introspection)

generates worthwhile ideas for designs it is worth pursuing, but the purpose is *not* "good" or "pristine" introspection as in (Hurlburt, 2011), but rather forwarding our human-like AI technology. So we may use introspective ideas to make designs, and later we may, if it behoves our purpose, abuse the introspection by using some introspectively-implausible parameters. Let a thousand flowers bloom (see point 7 in section 4.7).

7.3.4.2 Why this is more anthropic

AIF0 is more "human-like" or "anthropic", since humans do not have a pre-existing database of cases telling them that solution Y for situation X is correct, optimal, or "a solution" in some pre-determined sense. Humans struggle along in situations as they can (no "best"!) with the information they have.

The best response available to a human may be (objectively) quite bad, and moreover, once a "**bad habit**" of using a bad response has established itself in a human mind, it may be seen (subjectively) as "the best I have", even if an external observer can rightly judge the habit to be bad. A brief exposure to a better response to a situation may not cause an immediate overall switch-over to the better response, like any rational system would. By sheer "bad luck" or "pig headedness" the better response may be neglected (examples of this are in section 7.5.4).

Mistakes are made. This is a hallmark of non-rational, non-ideal AI.

7.3.4.3 Similarity

An important issue in many AI algorithms is the notion of **similarity**, often hidden inside a similarity function. The issue of how to judge the similarity of two inputs is easy only insofar as the input is *very* primitive, ideally discrete digital data. When the inputs are images with millions of multi-band pixels, or worse, videos, or "situations" (in the phenomenological sense) - then the issue of what is similar explodes into an impossible imbroglio. Considering our commitment to producing actual AI designs, we cannot just throw our hands up in despair à la Dreyfus (1979), but we have to, in each implementation, come up with some similarity function, just as CBR does. This gets more complex with the complexity of the examples, and always involves a wrenching sense of not doing justice to the real notion of similarity – but the software *must* be written. The similarity function, in principle, can involve another, entirely different AI

algorithm from the main one, such as one of the many derivatives of the neural net concept, "nearest neighbour" by Pythagorean distances in some vector space, a genetically-evolved similarity function, etc. We can also experiment with different approaches, as any technologists is allowed to do (see section 1.4.5).

One must recall that AIF0 is a simple example of the concept of AI based on introspection, and the resultant design is no more than a preliminary caricature of anthropic thought, a beginning of our project. We aim for phenomenologically more correct stuff, see the rest of this chapter.

7.4 AIF1

The purpose of this example is to show how introspection can be deepened, and more introspected mechanisms can be brought into software.

Introspection: I observe that all thoughts (including those about possible actions) have a time dimension, they do not appear as closed "cases" but as "sequences" over time. Once I commit in my mind that currently unfolding events are similar to some sequence of events in the past, I treat the whole sequence form the past as the "case" I am following, perhaps in a similar manner to AIF0.

Attempted approximation: I tried using an instance of the precious algorithm (AIF0) to determine the beginnings and ends of sequences, and use another instance (of AIF0) to select sequences and produce behaviour.

Having coded a version of this algorithm, it failed to produce behaviour better than noise (no useful learning). The algorithm that was supposed to find meaningful beginnings and ends to algorithms did not train meaningfully, probably (analysing retrospectively) because the idea that there are clearly defined beginnings and ends was wrong. Perhaps the success of AIF0 led me down the dangerous path of wanting my new "pet technique" to be *the* building block of future AI. This temptation to want all of intelligence to come out of a single idea is tempting, and several AI approaches have this "imperialist" view of the scope of use for their favoured idea: logical AI and neural nets spring to mind, as does the critique of this over-optimism verging on hubris (Dreyfus, 2012).

On the positive side, this is an example of iterative introspection. Having achieved a first success with AIF0, I went back and refined the introspection, adding a time dimension.

7.5 AIF2

The purposes of this example are:

- To further deepen the introspection.
- To demonstrate how complex introspections can be approximated by code, sometimes in an opportunistic way.
- To show how a failure (AIF1) need not imply a retreat, and that the solution could be in even more ambitious introspection (reminiscent of Dreyfus's demand for an even "*more* Heideggerian" AI (Dreyfus, 2007)).

Furthermore, this example demonstrates:

- Smooth acquisition of skills by interleaving multiple "cases", recalling Dreyfus & Dreyfus (1986).
- A possible concrete manifestation of Gadamerian AI as recommended by Winograd & Flores (1986), see sections 2.4, 2.5.

7.5.1 Introspection

The **introspective vision** became clear gradually – first I was displeased with the idea of discreet "sequences" - it seemed too constrained, too on-or-off. Then (while driving), I noticed how I was following *multiple sequences*, simultaneously – but could not quite put it into words. These sequences seemed to have to do with different aspects and eventualities of the driving, and of whatever was on my mind. However, this vision came initially with none of the below orderliness – all I could do to hold on to the vision was to point to the Beatles' song "Across the universe":

Words are flowing out

Like endless rain into a paper cup

They slither wildly as they slip away

Across the universe

Needless to say this is technologically useless. A more comprehensible **articulation** of this image emerged over the following few days, making the notions of “flowing” and “slipping” more concrete:

- The sequences (of AIF1) also known as “lines of thought” are not as on-or-off as in AIF1, but fade in and out without clear beginnings and ends.
- There are multiple such lines “being followed” simultaneously, to varying degrees.
- Actions are usually selected out of one of these sequences, again, aiming for the best future outcome, but often missing.

Over time and in different contexts the terms “episode”, “sequence”, “line/thread/train of thought” and just “line” were used, are equivalent, and represent a sequence of events in the past, being “followed” in the present, since it is similar to current events. Note that the sequences that are present but less dominant play the role of the “recesses of the mind”.

7.5.2 Introspective model:

In a mind, at any point, there are multiple different things “on one's mind”. These “things” (hereinafter “lines of thought” or just “**lines**”, sequences, etc.) are not static, but take the form of a sequence of events from memory, that are similar (or relevant, see below) to current unfolding events. Such “lines” are followed in the sense that the “current” part of the sequence advances in time as reality advances in time, like (when singing) we do not need to consciously “advance” in the lyrics and in the music when singing, but we just go with the flow of the memories of having heard or sung the song before. A 15-second phrase in the remembered song will be reproduced in a time not too unlike 15 seconds.

An action is selected out of the options presented by these lines of thought, as follows: if in line “a” I did A, and in line “b” I did B (and these are all the “lines” “on my mind” simultaneously), it is probably correct to assume I will do either A or B.

There is a preference to selecting an action that would be beneficial, so each line is given priority based on the anticipated reward occurring in the proximate time (the

anticipated future, or the yet to be replayed part of these remembered lines in the past)[22].

Of course this is still only a model, and many details are missing to recreate the original introspection. This is left to future work, see section 8.2. It is doubtful if we can ever fully model our introspections. But we also can't fully model real numbers in a computer – approximations are good enough for technology (see section 1.3.2).

The next step should be to find a formalization in software and data, that will create an approximation of simultaneous "lines" or "trains of thought", drifting in and out of some group of thoughts, perhaps called "consciousness".

7.5.3 Software design

The details of the design will be presented in an order that moves from some preliminary remarks, to the functionality visible in the introspective model, and then to the technicalities inherent in any design. These later technicalities are details of implementation that need to be addressed, like an integer's number of bits must be determined, and whether it is signed or unsigned; In this design, there are many such details.

7.5.3.1 Sequences in software

Recall that my introspections calls for "lines of thought" which "fade in in and out without clear beginnings and ends". This is not the common way of doing anything in computing. Usually we need a beginning in order to: 1. find the data, 2. not start before the data, over running something else, 3. have a starting point for some loop or process. We need an end for similar reasons: 1. Not to over-run some other data, 2. provide an end point where we can say our work is done. If we can achieve similar functions in AIF2 using different means, then we will have a reasonable design "without clear beginnings and ends".

Traditionally in programming, one would assume that a sequence would be represented by a formal array, a consecutive group of memory locations, or by a linked list. That means that a sequence has a clear beginning and end, and there is some pointer or index pointing to the "current" position. Consider printing out a string 40 characters long:

Assuming "a" is a character array representing a string, and "i" is an integer:

---

22 This is reminiscent of Husserl's "protension" (Beyer, 2015).

- for i=0 to 40
    - putchar (a[i])

Or consider printing a string which is a null-terminated sequence of bytes in memory, pointed to by “p”, “q” being another pointer:

- q = p;
- while (*q is not null)
    - print *q
    - q++

Consider further a linked list of characters:

- z = head_of_list
- while z is not null
    - print z.data
    - z = z.next

In both cases there are:

1. A data point marking the beginning: “a”, “p”, “head of list”
2. An index, traversing the data, “i”, “q”, “z”
3. an indicator of the end, “40”, the null character at the end of the last string, or the null pointer at the end of the linked list.

In AIF2, as we will see below, the beginning and end are abstracted away. There are separate mechanisms for “starting” or “ending” sequences. The only mechanism we need most of the time is just a pointer into the “current” moment in a sequence. This is a step away from the rationalistic view that we need strict control over beginnings and ends, and towards the subjective experience that all we have at any point is the now, with the various memories, thoughts and tunes going through our mind.

It so happens in AIF2 that the sequences are advanced by one time unit, represented by a “1” in the indexing of any area of the memory bank, but at the same moment the

memory bank is growing by one unit every iteration, so there is no danger that the sequence data type will “overflow” its end. We are now ready for the main design.

7.5.3.2 A novel data type

Here I introduce two closely related novel data types: a “sequence”, and a “table” of sequences. Sequences will usually be used in groups of a few at a time, such a group being called a “table”. The sequences are based on (and refer back to) a history database of all the history of the program's current run. The history database is not novel.

All the “personal history” of the AI system is saved, including inputs, outputs, and scores (from the outside world or micro-world). This array starting at 0 and ending “now” is represented as the “history database” in Illustration 7.3.

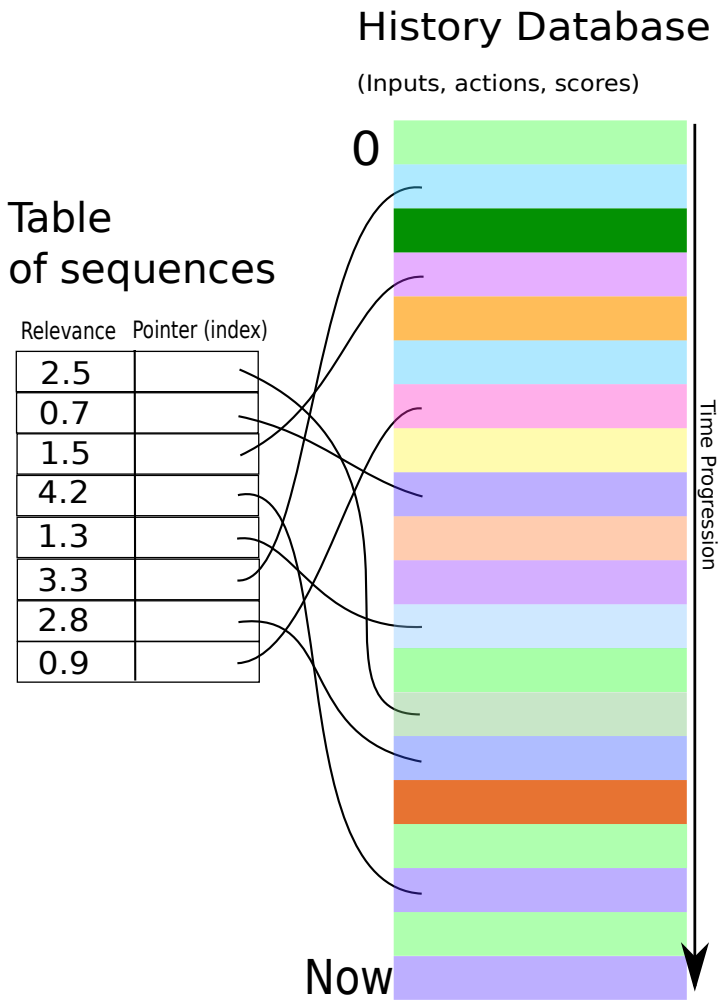


Illustration 7.3: AIF2 Data Types

The “sequence” data type is designed to represent a single train-of-thought, or line-of-thought, or a scenario that occurred in the past. Every “sequence” is minimally comprised of two scalar values: one is an **index** or pointer into the history database. This index points to a moment in time in the past that is analogous to the present moment in time. The **relevance score**, which is a (discounted) accumulating score of the similarity of the events in the sequence and the events as they are actually occurring in the situation facing the system. Every instance of a sequence represents a sequence of events (scenario in the past) that is considered to be similar to the present,

and therefore describes *a possible outcome* in the future, as seen from the current time. In Illustration 7.3, the similarity is illustrated using colours, representing the individual events. Note that the last two colour blocks are purple and green, incidentally very similar to the two blocks just prior, and hence the sequence following only two steps behind the present has a high relevance, “4.3”. The second most relevant sequence is in the very beginning of the history database, but the blue is not quite the same as the purple, hence the lower relevance of “3.3”, still high.

As these scenarios are inside a history database, populated by the past experiences of a specific run of the AI implementation, this represents the AI system’s own “personal” experience. Each sequence represents an option for action (the one taken at that point, or shortly thereafter, in the past), and also represents a possible outcome arising from the combination of the current situation and the action. Assuming that events occur in regular intervals, say of one second, then for every second that elapses in the “outside world” the index component of all sequences must be incremented, and the relevance score should also be updated based on the similarity of current events to the next event is the history database at the sequence's new “present”. Hence this is not a static data structure, but requires ongoing maintenance, to keep it “in the now”.

Typically a “table” would have 20-40 sequences in it. This data type is novel in that it is designed to imitate the way that multiple thoughts drift in and out of consciousness – therefore the sequences have no clear (abrupt) beginning or end (see section 7.5.3.1). In a sense the table as a whole represents the “consciousness”, and the relevance scores represent “how conscious” the system is of a particular train-of-thought or scenario from the past.

Technically, the sequences in the table each consists of a relevance score and an index pointing into a specific time in the past, that are part of a sequence of events that resembles current events. These pointers move forward in time in sync with the present time. The relative salience of each sequence to current events is measured by the “relevance” score (maintained next to the pointer in the table). This relevance is adjusted gradually, according to the similarity of current events to the events in the past sequence, so these sequences fade in and out in terms of their applicability to the current

situation.

Occasionally the least relevant sequences are discarded from the table, as their relevance score drops below a certain threshold. Once the table is depopulated (say below 10 sequences) it is replenished (this is a computationally expensive operation, see section 7.5.3.5).

Looking at Illustration 7.3, we can see 8 sequences in a table (on the left), each with a relevance score, and a pointer or index into a (greatly truncated) history database of all events (represented as coloured blocks) from the beginning of the run until the present moment. Assuming that the relevance score is overwhelmingly influenced by the last two events, you can see that the scores reflect how similar (or not) the last two blocks in a sequence are to the most recent two events in the chart, at the bottom of the coloured history database.

7.5.3.3 Decision process

A word is due about **relevance** (mentioned above) and **desirability**, which is the expected score over time in the future, based on a specific sequence of events in the past ("line").

These two key notions are similar, yet in a sense mirror-images of each other:

The "**relevance**" of a line is determined by how similar events *in the past* have been to *current events*. In a sense it is simply the (discounted) extension of the notion of "similarity" over time. Similarity is always already determined in the past – since we do not know the future we can not compare future events to the "future" of a line-of-thought. So "relevance" is in the past and plays a corresponding role to similarity, and is constituted from similarity. The difference is that relevance is stretched over time.

On the other hand, "**desirability**" is the expected value "promised" by each of the "lines", in the future. It is calculated from the "score" events in the "future" of each line, and is recalculated for the most relevant lines at each iteration. Desirability represents what the (naïvely) expected reward is in the case in which we end up repeating the same scenario. The AI system can "push" the present situation towards a repetition of a specific scenario by choosing the same actions that it took in that past relevant scenario.

In a sense, this is the whole point of the “intelligence” of this design: Try to repeat successful scenarios from the past. This is not dissimilar to CBR.

The decision process is as follows. For each action/output required:

- The top (say 10) most relevant lines are selected from the table.
- These lines are sorted by desirability, and the top (say 4) most desirable lines are further selected.
- A specific sequence (or random action) is chosen in a repeated round of coin-throwing, similar to AIF0. At a low probability a random action is chosen.
- The action taken in this past sequence is duplicated (or a random action).

The random element (like in AIF0) allows novel actions to be tried out within the normal process of responding to the environment, in contrast to many learning systems that have a distinct “learning” and “doing” phase.

Several possible future variants will be discussed in section 8.2.

7.5.3.4 More details of AIF2’s implementation

As we see from the above discussion of the design and from Illustration 7.3, there are many details missing from the picture. For now, let's assume that we have a sequence (or a table of them) already set up, and all we need to do is maintain these, and take decisions. Illustration 7.3 assumed that only about 2 events (the current and one previous) make the lion's share of the “relevance” of a sequence.

Every instant in time (the coloured blocks in Illustration 7.3) is an **event**. An event can either be an input from sensors, or an action that the AI algorithm outputted, or a score event.

At the base of this design is a **similarity function**, taking two events as arguments, that returning a value between zero and one that represent how similar they are, 1 for identical and 0 for completely dissimilar. The similarity function's exact implementation is not part of the specification of AIF2. This function can be hand-tailored or evolved using some learning algorithm.

Relevance is calculated as the previous relevance times some discounting factor, plus the current similarity minus 0.5. Note that relevance is not bounded in any range. If relevance drops below some threshold (another parameter of the design), then the sequence is dropped from the table. If a table is no longer sufficiently populated (drops below a certain amount of sequences) then it is repopulated, see below. This is the re-population threshold.

7.5.3.5 Dynamics of the scenario table

To populate a table of sequences (either initially or when the number of sequences drops below the re-population threshold), a scan is performed of the entire history database to find any consecutive events that would have high relevance, based on the currently recent N events, this N being the "**look_back**" parameter. In principle, if computational time were not an issue, such a scan should be done frequently, maybe with every input, so that any sequence that is relevant enough goes into the table, however, considering that sequences stay in the table for a while, it may be enough to scan for new pertinent sequences only occasionally. This also reflects the introspective observation that not all relevant scenarios from out past are always considered, we often forget things, in ways we do not understand. Matching this forgetfulness to the computational high price of repopulating the table is an opportunistic approximation – there is no good reason to think that these would match up, but we need to approximate in order to code (see section 6.3.5.2).

Because of the computational cost of scanning all the memory for populating the table, we can maintain a bigger table of scenarios than would be warranted by the introspection (say 20-40), and select only the top (say 10) "most relevant" scenarios every time we need to make a choice about an action. To choose an action from one of these 10 most relevant scenarios, we sort them by expected score ("**desirability**"), and like AIF0 select the best in 50% of the cases, and so on for 4 scenarios, with a possibility of a random action (see AIF0 in section 7.3).

Note that the active "scenarios" drift in and out of this "top 10" category, as they are selected from a bigger table. Also "scenarios" do not have predetermined beginnings or ends – when a situation (in the present) is similar enough to a situation from the past in the table it drifts into the high-relevance end of the table, and when a scenario drifts so

low in relevance it is silently dropped, but only a few iterations after it is no longer considered a “top 10” scenario. This gives this design some “softness” – mental events are rarely abrupt[23]. A sudden change in the environment would cause most if not all sequences to drop out of the table rapidly, leading to a re-population of the table that fits the new circumstances.

7.5.3.6 Initial conditions and decisions

Similar to AIF0, it is not a problem for a new instance of the system to act pretty much at random until it accumulates enough experience to draw from, so the initial conditions of the design are of secondary importance; we can assume that we have at least a minimal database of past events, including a record of inputs, actions and score events. This initial experience can be accumulated by simply letting the design flail around, randomly, for a while. We can also assume the table has been populated as per section 7.5.3.5. We must recall here that a complex system such as AIF2 may, like the “blue brain” project, often produce a non-functioning intelligence, or an intelligence uninterested in learning the tricks we present to it, see section 6.4, and the examples below.

7.5.3.7 Further Parameters

Many parameters are involved in this AI algorithm. Here is a partial list of issues and parameters that need to be tweaked is provided in order to demonstrate the complexity. In the experiments run so far many of these were assigned arbitrary numbers, and many (including the weights of the similarity function) were tuned using a genetic algorithm.

In terms of calculating “desirability” (see section 7.5.3.3):

- **look_ahead** – determines how many events in the "future" to consider when calculating desirability

In the context of choosing the sequence to use to guide current action, I have so far spoken of four iterations of coin-tossing at a 50% chance each. These numbers can vary from one run to another:

- **decisiveness** - the factor in the random decision, typically 0.5 .

23 Husserl's “extension” comes to mind.

- **max_cases** - how many times to try the coin tossing (and thus the chance of exploratory moves), typically 4.

The "decisiveness" parameter may change in time *during* run-time, following some function – for example it could become linearly higher.

- **var_decisiveness** - (boolean) use variable decisiveness
- **algebraic_form**- the equation for variable decisiveness, typically linear
- **vd_parameter_a**- the first parameter for calculating variable decisiveness
- **vd_parameter_b** - the second parameter for calculating variable decisiveness

In terms of table management, we also have:

- **table_length** - the maximum length of the table
- **think_about**- the maximum number of events to enter a non empty (and non full) table
- **init_relevance**- initial relevance of events that enter the table
- **max_relevance** - max relevance of events in the table
- **recent_prior** - (boolean) do recent events have priority (in the similarity function)

There are several more parameters. The point here is that the complexity of introspective algorithms is larger than those produced by mathematical models (Markov chains, GOFAI) but it is still manageable. As mentioned in section 3.2.1, the fact that human-like AI is difficult should not stop our exploration.

7.5.4 AIF2 Example runs

Some example runs are documented in three videos found in http://tinyurl.com/hycenh9

The same videos are found in the attached CD, as ".avi" files.

The three videos show the performance of AIF2 in a car-driving game, approximated from the work of Togelius, Lucas, & Nardi (2007). The purpose of these examples is to give *a taste* of what is achievable with AIF2, and more generally with introspection-based AI. Proper controlled evaluation and discussion of these results lies outside the

scope of this work, see section 1.2.1. Note that, as was expected, and as explained in section 6.4, relatively few runs learned at all. The examples below were selected for emphasising specific points, rather than for being typical.

The game in these examples consists of a “car” on a “track” that needs to race and accumulate as much distance in a given time. In the usual case (presented here) the car's driving software is entirely naïve at the beginning, and learns solely by getting a +1 score on moving in the right direction, +10 for passing a way-point at a predetermined proximity, while moving forwards (there are 8 of them, marked in blue, except the next one which is brown). The agent gets -10 points for colliding with a wall.

The car has 6 directional sensors (no diagonal sensors in the back) that feed in the distance to the nearest obstacle, and a directional (polar) feed pointing at the next way-point.

7.5.4.1 Learn 1

Discussing the experiment recoded in the video “**learn1.avi**” (see link on the top of section 7.5.4, or the attached CD): In this simulation there was a bug in the physics engine (leading to an undefined state) so occasionally the car-racing game is reset to its initial state, without resetting the AI software. The question of how skills are learnt best, in an ongoing engagement with the problem or with repetitive re-starts remains open.

Note that the car initially flails about, starts improving slowly, and by 1:35 manages to get beyond the “choking point” above the top right corner. By 2:30 some skill in not crashing begins to emerge, but mostly the car still crashes a lot. Around 3:30 it is completely lost. Around 4:30 it develops a (bad) habit of crashing into the wall repeatedly, but by 5:00 it rather suddenly starts to drive skilfully with no crashes for 65% of the round, recovers from a bad spell after a crash and completes several full rounds, with few crashes. The rounds complete at 5:26, 5:37, 5:48, and 5:58.

The learning process is reasonable, in terms of our expectations: it is cautious and experimental. The system has initial difficulties, and finds the top-right corner, which is tight, more difficult than the others. Once a full round is completed, the skills accumulated in traversing the beginning seem to kick in and serve to support the future circumnavigation of the track. The algorithm is behaving as expected.

7.5.4.2 Learn 2

Discussing the experiment recoded in the video "**learn2.avi**": The physics engine was fixed, so there are no "reset"s. Note that the track is a bit more difficult, with some more obstacles on the outer side of the track. After initially flailing about, going mainly backwards, at 0:14 the player gets into a strange mode of banging itself against a wall (in reverse). At about 1:00 it discovers that staying still is better than getting repeatedly penalized, and at about 1:05 it starts racing (quite wildly) towards the way-points. Other than another spate of banging against the wall at around 1:45, it learns to run around the track with less accidents, until the end of the video clip. It is interesting that "learn1" learned step-by step with great caution, while "learn2" spent much time being unproductive, but was significantly more reckless and "enthusiastic" in learning for the rest of the time.

The repetitive behaviour is interesting, not in being repetitive (that is something that most software does), but in being able to break away from the repetitive behaviour without any explicit intervention. Even after a pattern of behaviour establishes itself, the algorithm can break away from it. Here we see the benefit of the "random action". Once the programs gets out if this repetitive phase, its learning seems to display far less caution than "learn1".

7.5.4.3 Learn 3

Discussing the experiment recoded in the video "**learn3.avi**": It is a different track, and as usual the player flails about. However, in this case it develops a strange waltz-like strategy, of going around in arcs backwards, and only occasionally going forward just to collect the score at the way-point. It also (like "learn2" at 1:00) discovers at 1:35 that staying still is an option that is better than bumping into walls, but not as satisfying as getting more way-points. This "waltz" strategy works, but is suboptimal. In this it is like many habits of intelligent creatures.

7.5.5 Discussion of AIF2

Each of these runs develops **its own way** of doing things, like a "personality"[24]. These idiosyncratic development patterns are very encouraging in terms of producing

24 A highly qualitative observation is that the player or "car" in these videos seems (to several observers) to be rather "cute" - whatever that means. Exploring this question is outside the scope of this project.

anthropic AI, since humans are also diverse, non-optimal creatures (Ariely, 2009). Another point of similarity with the human condition is that there is no pre-knowlegde of the range of scores – AIF2 learns by preferring the better, not by aiming for the best.

One may raise the classical problem of "**credit assignment**" in AI – how does this design attribute the good or bad scores it may get from the environment to particular causes? This does not arise, since there are no cause-effect pairs, even statistically, like in a Markov chain data type (see section 7.6.1). The score events are simply registered as part of the history, and will sit there until that part of the history database becomes part of a relevant sequence, in which case the scores will be part of a desirability calculation, and hence will "motivate" the action taken at the time either positively or negatively.

-

AIF2 demonstrated a complex design, that deepened the introspection in the face of the failure of the previous design (AIF1). Using introspection as a basis for AI is not guaranteed to work, but in this case it did. This design is introspectively much richer than any other design encountered in the literature. Moreover, the introspected mechanisms as implemented in software present some surprisingly realistic behaviour, delivering some evidence for the anthropic approach (see section 3.2). Anthropic AI was defined as pre-programming only the basic intelligence, without any cultural commitments, and letting each instance of the program learn, or develop, its own "culture". The above runs demonstrated that a game can be learned from scratch, displaying a diverse range of responses to a similar environment. This diversity also speaks to the human-like character of this model.

Even though AIF2 is the last model in this document, it is not the end of the exploration. It should be seen as a basis for further development, see section 8.2

7.6 Consequences of the examples

7.6.1 AIF is more like CBR then like reinforcement learning

The AIF2 design can be seen as similar to reinforcement learning (RL), in that it navigates a situation using a closed list of actions, and takes feedback from the environment. It can also be seen as a similar to Case Based Reasoning (CBR, see

section 7.2) since it uses stored episodes as “solutions” for future situations. The RL paradigm includes the notion of a Markov chain, and of a closed universe of states that can (at least in principle) be explored exhaustively. CBR on the other hand is more open ended, and more modest, in that it does not aim at an overall solution (a fully understood Markov chain) but at “the best we have”. In this sense it is closer to the AIF family, which being introspective does not aim (even theoretically) at infallibility. Moreover, the AIF family of designs has as its main data store a historical database, like CBR’s store of “cases”, and not some statistical summary of weights or probabilities like RL. AIF0 even retains the separation into discreet “cases”.

7.6.2 The “sequence” data type

Probably the most interesting technical contribution that AIF2 makes is the introduction of the “sequence” data type as a building block for AI systems, see section 7.5.3. This allows the representation of multiple fading trains of thought that originate from the memory of past experiences.

Even if AI practitioners were to ignore all the arguments presented in this thesis, and will not consciously seek to use introspection, the mere introduction of such subjectively-informed designs into the discourse will broaden the field of “allowed” or “respectable” discussion. Considering how infrequently non-philosophers read philosophy, probably that is the most that can be hoped for in terms of any impact of this work on the technology debates and marketplace, at least in the short term.

7.6.3 Dynamic symbols

Consider AIF2 as a model of the working of the mind. It can be viewed as matching events from the past to the present, if you will “interpreting” the present using the past as the source of possible scenarios with which to understand the current situation. Each AIF2 run creates its own history database which it uses to interpret future events.

In classic AI designs, the symbol-system in terms of which the world is construed is pre-determined by the programmer or knowledge-engineer using a fixed vocabulary. The Classic AI system does not “grope around” looking for a vocabulary through which to construe a situation. Neural nets (in the learning phase) do “grope around” and are adaptable in unpredictable ways, but we cannot see (at least usually) a signifier-signified relationship at all. In AIF2, conversely, the very *terms* of understanding are a

product of the lifetime of the system itself. AIF2 understands the current situation in terms of its previous memories. A past memory that is used often can be construed (by us, as observers) as serving as a symbol, a concept or a metaphor “describing” or “interpreting” the present.

It is as if Classic AI's version of the human has been given a language by some supernatural being, and the designed “mind” is confined to use only that language while AIF2's “mind” evolves its own concepts to deal with any situations that it encounters with any regularity. This development of the mind may even occur socially, in future settings. This is reminiscent of COG (see section 3.2.6).

This has technical and philosophical consequences:

Technically:

- Since there is a symbolic relationship, there is an “understanding X in terms of Y”, one can debug such a system more intuitively than a neural net, without committing to the rigidity of a predetermined set of symbols.
- A system that has both plasticity and sentence-like structures can adapt to cultural factors that are particular to the environment of every particular run, even in a temporary manner. Classic systems have little plasticity, and Neural nets have nothing that resembles sentences or structures that could accommodate the cultural transmission of habits, see also section 5.2.
- A more sophisticated system based on this approach also would allow for reflection on its own practices: introspection by the AI rather than by the developer (contrast with section 3.1.2)

Philosophically, this notion of viewing past sequences as representations of sorts can be an interesting input to the debate on representations (Shanon, 2008): In a new sense, AIF2 harvests it's own symbols, turning past events into symbols and past sequences of events into representations for understanding the future and speculating about it (see also section 8.4.4).

7.6.4 How AIF2 is Gadamerian

AIF2, as presented above (section 7.5), reflects many of the characteristics that

Winograd & Flores (1986) view as interesting in Gadamer's hermeneutics. Recall the introduction to hermeneutics given in sections 2.4.1 and 2.5. This avenue of research (though explicitly suggested in (Winograd & Flores, 1986)) seems to have been neglected in the AI literature.

AIF2 does little but match previous episodes from memory to the current evolving situation. Even doing just that can be seen as interpretation of a crude type, a bit like CBR interprets the situation as being similar-enough to a past “case”. But AIF2 goes further, and allows the current situation to be matched with more than one sequence, and thus be interpreted and reacted to using a *blend* of previously experienced scenarios.

Recall the name of Gadamer's magnum opus (2004) - “Truth and Method” - Truth stands for the brute facts of the text (or the “sense data”), and Method is all the wisdom, methodology and experience the reader brings to bear. Every person's “method” is a result of their own life experience, education and other memories – ultimately these are all stored in the individual's memory (discounting any Jungian-type “collective unconscious”). So according to Gadamer, all interpretation is done using the past memories of the individual – and the AIF family of designs provide an underlying mechanism for implementing such a system. Recall also that Gadamer viewed the act or interpretation as a “merger of horizons” - the horizon of “truth” - sense data, and the internal horizon(s) – of “method” or memory. Gadamer also stressed the unavoidable existence of “prejudices” - and we can see even in AIF0 (section 7.3) that once a “habit of thought” is established it is difficult to dislodge.

7.7 Examples of introspection being used for AI design: summary

I have argued for introspection-based AI, and showed some designs as examples, and also showed that the more sophisticated example design (AIF2) may have some interesting potential. Since this thesis is in philosophy of technology, the ultimate purpose is to make a contribution to technology.

Note also that this entire chapter of examples is nothing more than a down-payment towards further development and a more thorough evaluation of such algorithms, which I outlined as the “third volume” of this project, in section 1.2.1. The main body of this document is the middle volume in the outlined trilogy.

8 Conclusion & possible consequences

Table of Contents

8 Conclusion & possible consequences....194
8.1 Conclusion....195
8.2 Future technical work....198
8.3 Possible consequences for cognitive science....200
8.3.1 Models for scientific psychology....200
8.3.2 A response to Dreyfus's critique of AI....201
8.3.3 Natural language processing....201
8.3.4 Cognitive models....202
8.4 “Underpinning” models in philosophy....202
8.4.1 Wittgenstein's aspects....203
8.4.2 Gadamer....204
8.4.3 Dreyfus's demands from AI....204
8.4.4 Wheeler's action-oriented representations....205
8.4.5 Adhyasa / superimposition....206
8.5 Open Questions....206
8.5.1 Dilthey vs Gadamer....206
8.5.2 Further unexplored terrain....207

This last chapter concludes (section 8.1) and discusses the possible consequences of this work:

- For AI Practitioners, future possible extensions of the AIF family of algorithms are presented in section 8.2.
- For Cognitive scientists, the possible impact of this research is discussed in section 8.3.
- The manner that introspective ideas developed here may serve to “underpin” some ideas in philosophy is presented in section 8.4.
- Some more general outstanding questions are left for section 8.5

In discussing *possible* consequences (future work), the arguments are less conclusive than in the main part of the thesis.

8.1 Conclusion

This thesis argued, in the field of *technological* AI that **introspection is recommended for anthropic AI**. It explored the conceptual space between phenomenology and AI, mainly classic AI.

A double crisis exists in AI: there is a dearth of new conceptual frameworks (section 1.1), and there is a neglect of the actual complexities and paradoxes of human thought (as revealed subjectively) in favour of a rationalistic viewpoint (section 2.4).

The project of this thesis was to mitigate the faults of excessive rationalism using the subjective (standing with Dreyfus and Winograd & Flores, against Simon and his followers). This turn to subjectivity is done while still being committed to producing concrete working software (standing with Simon and the mainstream AI community, against Dreyfus).

Human-like AI was discussed as distinct from the ideal/rational type following S. Russell & Norvig (2013) (my section 3.2.1). Human-like AI is required for applications where smooth interaction between unskilled people and robots is key (section 3.2.2). Human-like AI has so far received far less attention that ideal/rational AI, both in research and in education. Technological AI was distinguished from scientific AI (section 1.5) - this thesis focused on human-like AI as a technology.

Within the context of the search for human-like AI, **Anthropic AI** was defined as emulating the fundamental intelligence inherent in humans that makes the learning and acquisition of culture possible. This was contrasted with the “western, modern, well-trained and adult” intelligence that is so often sought after in AI, but is a contingent fact which is only true of our particular current culture (section 3.2.4). COG and CYC were recognized as earlier efforts that initially constructed a fundamental intelligence, and then aimed to acquire the necessary skills and/or knowledge by a learning process (section 3.2.6). Scientific models of the human mind (available for technological implementation) provide only models that are “too high” (like logic and cognitive simulation) or “too low” (like neural nets) for the purposes of anthropic AI.

Starting the work of rehabilitating introspection for the purposes of AI, **subjectivity** was shown to be a valid angle of research for AI (section 3.3.1), and some examples of cognitive science touching on subjectivity were presented (section 3.3.1.4). Within the

subjective realm, the phenomenological critique led by Dreyfus was surveyed (section 2.3), and critiqued for being only negative, not producing any concrete software designs.

Perhaps the most subjective approach of all, **Introspection** was discussed and shown to be suspect not only for reasons given by J. B. Watson (1913), but also in-principle: It is impossible to make neutral, interpretation-free observations in the natural sciences, so we have no reason to expect the situation to be any better in the subjective realm (section 3.3.3.4).

Regardless of the above, cognitive science treats introspection as being an **illegitimate** method (Nisbett & Wilson, 1977; J. B. Watson, 1913, 1920). Bringing in the distinction between the context of discovery and the context of justification, *any source of ideas* in science is legitimate, so introspection is rehabilitated as a legitimate source of discovery, of ideas, for science (section 4.2.4). Moreover, the level of truth required for *technology* is significantly lower than for science (sections 1.4, 4.2.5) so introspection (for AI) has been fully rehabilitated from the traditional 20th century view that it is "wrong", "disallowed", or illegitimate for some other reason. Somewhat strangely, AI researchers *have* used introspection, not least Herbert Simon (section 4.3). However, they have not used introspection full-bloodedly – they most often tend to expunge any reference to introspection from peer-reviewed publications, and deal with it sparingly and shyly (sections 4.3, 4.6). Sherry Turkle's description of how the field of AI relates to introspection (section 4.3.1) came nearest to my analysis, but it remains sociological and factual rather than analytical. Phil Agre also gave an outline quite close to mine, but shied away from introspection as such, eventually trying to create Heideggerian AI, but "*by his deictic representations, Agre objectified*" the ready-to-hand, thereby missing the phenomenological point (Dreyfus, 2007). Agre did not do introspection for AI as I propose (see section 4.3.2).

Chapter 5 showed that introspection is a positively **plausible** basis for AI, since it is used reliably in education. In teaching skills, in order to generate their narrative, instructors either recall the narrative used to instruct themselves years before (unlikely), or generate a narrative by self-observation. When the skill being taught is a mental skill, this involves *mental* self observation – introspection. The success civilisations have in

transmitting mental skills from one generation to another serves as testimony that this kind of introspection works, i.e. introspection *is efficacious* in transmitting skills from one human to another, and therefore introspective reports contain some sort of information *that may well be efficacious* in replicating human mental skills in software.

In the examples (chapter 7), the use of introspection as a basis for designing AI software was demonstrated, and the new possibilities allowed by the rehabilitation of introspection for the development of AI designs were shown. Previously, insofar as introspection was used at all in AI designs, it was used as a basis for a single mechanism, which was later integrated into the mainstream mathematical framework of the discipline. The examples given (in chapter 7) showed that *multiple* mechanisms can usefully be adapted from human introspective reports into an AI system. Sometimes where one or two mechanisms fail to produce a useful result, going for *more* introspection rather than less allows for the creation of a working system. This was impossible while introspection was minimized, treated as somehow illegitimate in AI research contexts.

An advanced example (AIF2, in section 7.5) was described that brings forth a novel data type that not only represents past episodes, but in a sense allows them to be re-enacted in the mind, in sync with current experience. Multiple such trains of thought fade in and out of significance. This was analysed (section 7.6.3) in terms of representations, a central area of debate in cognition, and the notion of "dynamic symbols" was defined. These symbols are used in systems that understand the world through a vocabulary that is *not pre-defined.* This was shown to be a concept that fits in with the Gadamerian world-view (see sections 7.6.4, 8.4.2), thereby providing a concrete manifestation of one of Winograd & Flores's desiderata. This can be seen as a step towards Heideggerian AI (see section 2.5.2), since Heidegger himself viewed Gadamer's work as a detailing-out of his own work in hermeneutics.

The subsequent sections of this chapter will examine the impact this work could have on AI, cognitive science and philosophy.

The Appendix is US Patent No. 8,660,670, detailing the engineering novelty of AIF2.

8.2 Future technical work

Beyond AIF2 as presented in section 7.5, the following variants could be of future interest:

1. The internal parameters of the algorithm may vary over time. For example, this has already been implemented regarding the weights of the randomised selection. This can make the algorithm more "decisive" or "conservative" - i.e. more likely to select the better episodes over time. This is based on the introspective observation that with skill less experimentation is done, and the external observation that people (and other mammals) become more conservative as they age.

2. The set of **actions** (available in the environment) may vary over time, or as a result of developments within the game/environment. This is planned to accommodate the discovery over time that some actions are not available, or the stumbling on to new options. This further moves away from classic AI's Markovian assumption that the group of possible actions is an (often small) finite set.

3. The **score** (given by the environment) is currently a simple number indicating an overall assessment of the outcome, and could be comprised of a few separate components, for example, indicative of progress towards low-level and high-level goals, or short-term and long-term goals. This is an attempt to deal with the existence of several goals, perhaps on different scales, simultaneously.

4. Even in the case of a random action, various actions correlated to particularly undesirable outcomes may be excluded from the range of options ("**Panic** mode"). This is based on the observation that we may be adventurous only within certain bounds, and not only reward-seeking, but also catastrophe-averse.

5. The **history database** may be entirely "real" (derived from the current run) or may be a "manufactured history" including "transplanted experience" from another run, and/or a manually encoded history, to provide the algorithm with a starting baseline of experience. This is to allow for pre-taught robots, or for Chomskyan (or Plato's Menon) pre-known skills[25]. The transplanted history could also be derived from some

25 This is a bit reminiscent of the "young earth" argument trying to reconcile the geological record of life existing for millions of years with the biblical story that makes the earth under 6,000 years old. This theological exercise points out that God could create the world with all this record already embedded in it. Likewise, an AIF application instance could have a past memory that did not actually

process, such as crossing between two “parents”, providing a new medium for genetic ideas.

The idea of starting afresh while being “convinced” that there is a lot of history is reminiscent of how computer systems resume after being “hibernated” (as opposed to “suspended”).

6. **Pruning** of the historical database can be performed on the basis of discarding the oldest data (suitable for rapidly changing environments), or on the basis of other criteria such as discarding historical episodes which resulted in mediocre desirability of the outcomes. These (mediocre) memories would not be particularly useful either for re-use or for avoiding past mistakes. This is based on the introspection or observation that we forget mainly the mundane.

7. **Federating** – This algorithm may be used in multiple instances, sharing the same time-line and history database. This could be useful for example to control different time-resolutions, or to have some “controller” select which of several machines implementing different skills/strategies should be used at any time. This is also an attempt to deal with the existence of several-scale goals simultaneously, and other possible complexities.

8. The parameters of the **similarity** function may be tuned by a genetic algorithm. This is to address the fact that we have no special insight (introspective or otherwise) into the well-known problem of defining similarity.

9. The other (“**technical**”) parameters of the algorithm may also be tuned genetically. Examples of such parameters are table size, thresholds and the parameters for randomised selection. This is not introspectively motivated, but comes to tune the many arbitrary decisions made during the process of approximating from the introspection towards a formal algorithm.

10. The time parameters could be made more continuous, in a sense using real numbers as time indexes. This would move further away from CBR, in that time becomes less atomic. It would also require the similarity function to compare not atomic events, but “moments” that would perhaps have some duration[26].

---

occur within the run, but was preloaded.

26 This again is reminiscent of Husserl’s extension and protension.

11. So far, sequences are recalled in “real time”, the sequences being advanced by one time unit every for every time unit elapsing in the external world. Allowing some flexibility will allow for slowing down or speeding up the memories. This will allow the AI to experiment with applying skills at varying rates.

12. In a sense, in AIF2 the sequences of memory are themselves being experienced, in that they “play out” in sync with the unfolding (real) events, and move in time. But in deeper sense, this mere following-along is part of the interpretation, but not part of the experience being interpreted, in that the recalled sequences never move into the “current sense data” that is compared backwards to previous sequences in memory. One could envisage such a system, where we explicitly make the contents of past sequences being recalled part of the “present sense data”. This would be a most interesting area of research, since this opens up the possibility of an AI system recalling “that it had such a combination of thoughts before” and suchlike. This may also be the beginnings of self-reflection, perhaps machine-consciousness (Gamez, 2008), and perhaps even introspection-by-the-AI, see section 3.1.2.

8.3 Possible consequences for cognitive science

8.3.1 Models for scientific psychology

An important distinction used to build this thesis can now be relaxed a bit, in order to show more potential utility from this work. Parts of chapter 1 were spent making a clear distinction between technology and science, and some of chapter 4 (especially section 4.2.4) presented the distinction between the context of discovery and the context of justification in science. I claimed that introspection was a legitimate source of ideas because it was only used in the context of discovery, and moreover it was only being used (in the main body of this thesis) for technology. We can now relax this a bit and discuss the interaction between these. In a sense we “*must throw away the ladder*” (Wittgenstein, 2001b, sec. 6.54).

Recall that:

1. All ideas are allowed in the context of discovery *even* in science, that is where this distinction originated from.

2. Psychology already uses some AI programs as models, or sketch-theories about

how cognition works (Sun, 2008), as Simon predicted that "...*theories in psychology will take the form of computer programs*" (Simon & Newell, 1958).

AIF2 (and other introspection-based AI designs) may be used as theoretical tools in psychology. In being introspective, these theories perhaps would have a better chance (than for example neural nets) to bridge the cognitive and personal-analytical branches of psychology.

8.3.2 A response to Dreyfus's critique of AI

AIF2 (section 7.5) uses dynamic symbols (section 7.6.3) "mined" so to speak, out of the input stream, to signify / interpret other events in the input stream. Unlike Simon's classic AI, there are no pre-fixed symbols, but nonetheless there is a certain notion of representation (see also section 8.4.4).

Classic AI, insofar as it used introspection, used it both sparingly (simulating one mental mechanism at a time) and in an idealised manner. However, they did produce concrete working systems.

Dreyfus seems to get carried away with the "what computers can't do" slogan. His argument would have been better served by being called "What can't be formalized". But heavily informal systems, such as global weather, can be simulated by computer to any precision we choose. His argument that a 100% emulation of a human mind in a computer is impossible should not stop us approximating. My thesis agrees with Dreyfus that if one is looking to get to the moon one should stop climbing the nearest tree, and get into a warm room for some deeper planning. However, Dreyfus's discussion doesn't elevate us towards the moon even the few meters that the tree would, since no working software is produced by his approach at all (Dreyfus, 2007). All Dreyfus does is show why Platonism/intellectualism will not get us there. He stops there, without proposing any concrete alternative that is better than that first tree.

This is why I attach working examples to my argument. Dreyfus's founding contribution to philosophy-of-AI is not being belittled, but it is time to move forward to a more positive contribution, and show what *can* be done, not just what *can't* be done.

8.3.3 Natural language processing

An interesting application of the AIF family could be in natural language processing.

Again, no strong claim is being made at this speculative stage that my suggestion would be better than any other – only that this is worth exploring.

A more advanced version of AIF, where there would be (at least) two AIF2-type mechanisms cooperating could be used to produce **language** – with one engine following structure (including syntax), producing (more-or-less) grammatical sentences, while another instance deals with the content, bringing up the relevant semantic fields – and somehow cooperating in producing the eventual sentence. Such sentences would hopefully be *mostly* but not completely grammatical, mostly but not completely on-topic – quite like human speech.

Conversely, in terms of understanding natural language, a more advanced edition of the AIF designs could also be used to understand different aspects concurrently – for example following the various grammatical rules and customs on the one hand and helping build a mental "picture" using the meaning of words by placing the various recalled meanings in the correct relations to each other (as the previously-learned grammar dictates). This would require having some "Cartesian theatre" or "imagination" able to draw mental pictures, and the ability to react to these pictures in relation to the external reality.

8.3.4 Cognitive models

AIF2 is entirely consistent, as an example for introspection-based AI, with some of the most popular views in cognitive science. The "**predictive brain**" concept (Clark, 2013) argues that the main (if not only) function of the brain is to predict the environment and to manoeuvre the person's body in such a way as to both minimize prediction error and produce the best possible results.

8.4 "Underpinning" models in philosophy

In this section I will show how AIF2's "dynamic symbols" (section 7.6.3) can also be seen as underpinning various ideas in cognition and philosophy. I use the term "underpinning" specifically to make a different point than "support". One supports a world-view with arguments, showing why one should accept or believe in some position. When I say "underpinning" I mean that one could construct a software model that would operate in a way similar to the intellectual idea being examined. In a sense,

this allows a philosophical model to “come to life” *in silico*. In another sense, “underpinning” is a software-based experiment, somewhat similar to a thought experiment. Like the thought experiment device, it *may* lend support to a model, but that would require clear argumentation. I will here argue that AIF2 or subsequent members of this family of designs can be used to underpin several philosophical concepts:

- Wittgenstein’s “aspects”
- Gadamer’s “prejudices”, or “method”
- Dreyfus's demands from AI
- Wheeler's “action oriented representations”
- Indian philosophy's *adhyasa*

The sections below are dedicated to showing how, perhaps with some imaginative license allowing for future research, the above philosophical concepts may be underpinned by an AIF2-like mind. Again, none of this is to claim that AIF2 or its derivatives are true, correct, or even superior to any other designs, past or future. The sole purpose is to show how in a wide range of senses introspection-based designs can be useful.

8.4.1 Wittgenstein's aspects

In (Wittgenstein, 2001a, p. 166) we are introduced to the duck-rabbit drawing. When presented with this picture, people usually either persistently say that it is a rabbit or a duck, or that they see that both can be seen in the same picture. Most people after discussing the dual nature of the picture can see the duality.

Many AI programs that are logic or statistics based (say an expert system) would assign a higher probability to one or the other interpretations, by some function. Perhaps indeed the picture in (Wittgenstein, 2001a, p. 165) is by some objective measure more a duck than a rabbit, but that is not how it is *for us* humans. AIF2, on the other hand, would either have past experiences of seeing the picture as a rabbit or as a duck, and would repeat that habit, probably (just as a human) until shown the other option, and would then alternate between the two haphazardly. The very structure of AIF2 is similar to Wittgenstein's “seeing as” - it acts in any specific situation X based on one of a

number of similar situations in the past. In that sense, AIF2 sees the current situation as equivalent to that past scenario.

8.4.2 Gadamer

Gadamer's view of hermeneutics is that we use our memories of the past (which form our "prejudices" or "inner horizon" or "method") to interpret the present (termed "truth" or "outer horizon"). We do not interpret the present using only one memory at a time as in CBR, but use a blend of the memories available to us. AIF2 is to the best of my knowledge the first attempt to bring to bear on AI decisions such a time-based blend. For more details of how AIF2 is Gadamerian see section 7.6.4.

8.4.3 Dreyfus's demands from AI

As we saw in section 2.3 Dreyfus's critique of AI is broad, and unrelenting (Dreyfus, 2007). Dreyfus also teaches us the dangers of "first step" fallacies (2012), so I cannot even claim that AIF2 or the whole introspective approach is a good "first step" towards appeasing his demands. What I can claim, that it is *a* step, perhaps "up a tree", but at least up a taller tree than existing AI, at least in the following sense.

In (Dreyfus, 1979, p. 253), within a complex and broad context, Dreyfus outlines one specific element of what a more human way of doing things might be, in one type of skill acquisition:

> *Now suppose that, in this random thrashing about, I happen to touch something, and that satisfies a need to cope with things…. I can* ***repeat whatever I did****, this time* ***in order to*** *touch something…. This is presumably the way skills are built up.* (stress in the original).

AIF2 does nothing if not "thrashing about" and later, in encountering similar circumstances recalling the better and worse outcomes that were obtained in the past, and giving a higher chance to "repeating" the beneficial action in the future. Like Dreyfus's description, there is no sharp distinction between a "learning phase" and a "using" of some formal stored "knowledge". AIF2 therefore implements discovering skills and learning to use them in a more Dreyfus-like way than anything done before in AI (as far as I could discover).

If that were the entirety of Dreyfus's world-view, one might be tempted to say that it is similar to reinforcement learning, but in Dreyfus's critique it is part of a broad

Heideggerian world-view. The view presented in this thesis is quite Gadamerian (see section 7.6.4), and again, Gadamer's hermeneutics is only one element of the Heideggerian world-view. So yes, perhaps even every *element* in our AI could be compared to existing systems, but the power of AIF2 is specifically in the blending, and the power of this thesis is in showing how we can use introspection to generate *many* AI designs.

In showing that AIF2 makes *a step* towards answering Dreyfus's concerns, I have shown that *one example* of introspective AI made one step, and hence one could plausibly expect further engagement with introspective AI to make some more positive steps. Not only is AIF2 interesting to this debate *per se*, introspection-based AI is a *generator* of interesting AI designs, and hence "recommended for developing anthropic AI" as per my main thesis, and a substantive response to much if not all of Dreyfus's critique.

8.4.4 Wheeler's action-oriented representations

Wheeler (2005, pp. 195–196) recalls that orthodox representation theory (such as in classic AI) calls for representations that are "*essentially objective, context independent, action-neutral, stored descriptions of the environment*". He contrasts that with Brooks's "*situated robots*" that have no representations as such (recall Brooks's paper was entitled "*Intelligence **without representation***" (1991)).

Wheeler mentions Clark's (1998, pp. 47–51) action-oriented representations as being half-way between mirroring the world and prescribing action. However, Wheeler critiques this "mid-way" approach: the living organism has no interest in representing the world objectively-as-it-is, but only in terms of (ego-centric) actions. The representations of actions-and-outcomes do indeed reflect something of the world, but that "purely academic" part is of no interest to the living creature - "*it is by adaptively mediating between sensing and movement that such inner structures earn their keep*" (Wheeler, 2005, p. 197). Wheeler also points out that these structures represent knowledge-how, not knowledge-that (see section 3.2.7). Note also that such a direct association between sensing and moving, without much if any intellectual content, is also fundamental to O'Regan's (2011) sensory-motor theories.
AIF2's "sequences" and "dynamic symbols" (sections 7.5, 7.6.3) are good candidates for the role of Wheeler's action-oriented representations, at least as far as technological AI is concerned. They are used to predict possible consequences if certain actions are taken, based on the system's own past experiences. Arguably that can be found even in CBR, not only in the AIF family. Unlike CBR, in AIF2 the representations/sequences can fade in and out and "cooperate" or "blend" – they are "softer" and allow for more

subtle skills, and skill combinations.

8.4.5 Adhyasa / superimposition

Adhyasa is the Sanskrit[27] term for “superimposition” - it refers to people seeing in a situation some thing, distinction or circumstance that is not in the actual input. An example could be meeting a young person on campus, without conversation or prior acquaintance, and seeing them from a distance, perhaps carrying a computer, as an undergraduate. This idea of the bureaucratic definition “undergraduate” is added by the observer. Moreover, if you have just imagined the described situation vividly, you most probably imagined either a young woman or a young man, and they had some other specific features, like hair, clothing, etc. Not one of these details were in my description. All these details are added by the listener, “superimposed” on the basic situation, not really there. Similarly all social constructs such as belonging to a nationality etc. are not part of the objective reality, but are superimposed by our understanding.

Again, like in the case of Wittgenstein's seeing-as, the case of adhyasa is also similar to AIF2. The AIF2 agent is reacting according to the *predicted* consequences of its actions, predicated according to past experience. It “sees” probable consequences before they arise. They may well never arise, since the situation at hand may be entirely different from the memories of the AIF system. Again, no claim is being made that AIF is the only AI concept that could produce such behaviour. In the case of adhyasa the phenomenon of “over-fitting” in machine learning comes to mind.

8.5 Open Questions

This thesis has been an exercise in inter-disciplinary work. I have already mentioned several areas for further research, much of which could be done by experts in the relevant specific fields. I am left with some observations and worries that do not fit neatly in any one discipline:

8.5.1 Dilthey vs Gadamer

In section 7.6.4 I argued that AIF2 can be seen as a start in the pursuit of the long-awaited Gadamerian AI. However, I did not advocate for taking phenomenological or hermeneutic texts as “gospel truth”. I argued for personal introspection by the AI

27 Sanskrit is to India what Latin and Greek are to Europe.

developer as a way to get at some mechanisms of the human mind in order to emulate them in software. In a sense, I am promoting a Gadamerian approach for the content of the AI software, following Winograd & Flores (1986), but I promote a Dilthey-like approach to obtaining the ideas for building the software. Recall (from section 2.5) that Dilthey called for a balance (in the person trying to understand some historical text) between "living experience" and critical thinking. For Dilthey "living experience" meant that one should allow oneself to imagine what it would be like to be, say, Caligula in a certain situation, and try to use one's own reconstruction, like an actor's, to understand the dynamics of the situation. He then called for critical thinking to restrict those flights of fancy, and make sure that they make sense in the historical context, and in whatever other constraints are known to apply in the historical situation (Ramberg & Gjesdal, 2014).

In this thesis I suggest that one should use introspection in order to create in software models of how we humans may think. I do not restrict this to "correct" introspection – I explicitly allow "bad", speculative introspection in (see section 3.3.3.5). This can be seen as the equivalent of Dilthey's "living experience". My version of the restraining critical historical thinking is the demand that specific operational systems must be produced, run, and evaluated as a technology.

8.5.2 Further unexplored terrain

Two unexplored areas in cognitive science have been mentioned, that cry out for exploration – perhaps even more urgently for my specific interest in anthropic technological AI.

1. First, Anthropology has been pointed out by Boden (2008, Chapter 8) as "the missing discipline" in cognitive science. My term "anthropic AI" only scratches the surface, in that I am interested in humans-as-such rather than in "western, modern, well-trained adults". However there may well be some anthropological literature about cognition that has not been explored. A non-European phenomenology would be fascinating, which brings up the next point.
2. The literature on the structure of the mind in Sanskrit is vast. This is difficult to explore since beyond being mostly untranslated, these Indian traditions do not separate psychology, metaphysics and religion to any degree that even begins to

satisfy our western, modern expectations. Again, something being difficult is no argument for expecting it *a-priori* to be wrong. Specifically for our interest in AI specifically and cognitive science in general the Buddhist tradition has an obsession with enumerating lists of mental components of various mental mechanisms, if "mechanisms" is indeed the correct term.

Final words

This thesis argued that "**introspection is recommended for development of anthropic AI**". This should perhaps be the beginning of a retreat from science's domination over other areas of thought, specifically our thinking about human nature, but also our thinking about technology.

In cognitive science, as Searle (1992, p. 115) put it, we should (supposedly) "*carve off and eliminate the subjective experience*". I call for a moderation of that programme. Little be it for me to make direct recommendations about cognitive *science* in general – but in the field of human-like AI as a technology, this attitude has surely outlived its usefulness. Instead we should return to an age-old piece of wisdom: "*It is the mark of an educated man to look for precision in each class of things just so far as the nature of the subject admits*" (Aristotle, 2009). By all means we should render unto science that which belong to science, but let's move on with the technology. We needn't wait for scientific precision.

9 Appendix - US Patent No. 8,660,670

This is presented as further evidence of the novelty of AIF2 (see section 7.5)

US008660670B2

(12) **United States Patent**
Freed

(10) **Patent No.: US 8,660,670 B2**
(45) **Date of Patent: Feb. 25, 2014**

(54) **CONTROLLER WITH ARTIFICIAL INTELLIGENCE BASED ON SELECTION FROM EPISODIC MEMORY AND CORRESPONDING METHODS**

(76) Inventor: **Sam Freed**, Jerusalem (IL)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 408 days.

(21) Appl. No.: **12/876,266**

(22) Filed: **Sep. 7, 2010**

(65) **Prior Publication Data**

US 2011/0060425 A1 Mar. 10, 2011

Related U.S. Application Data

(60) Provisional application No. 61/241,044, filed on Sep. 10, 2009.

(51) **Int. Cl.**
G05B 13/02 (2006.01)

(52) **U.S. Cl.**
USPC **700/49**; 700/108

(58) **Field of Classification Search**
USPC 700/2, 9, 20, 28, 29, 30, 33, 34, 47–51, 700/108–110
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

6,490,519 B1 * 12/2002 Lapidot et al. 701/117
7,103,460 B1 * 9/2006 Breed 701/32.9

OTHER PUBLICATIONS

Ros R et al :"A case-based approach for coordinated action selection in robot soccer", Artificail Intelligence, Elsevier Science Publisher B.V.,Amsterdam, NL, vol. 173, No. 9-10, Jun. 1, 2009, pp. 1014-1039.

Mirza N A et al :"Developing social action capabilities in a humanoid robot using an interaction history architecture" Humanoid Robots, 2008, Humanoids 2008, 8th IEEE-RAS International Conference on, IEEE, Piscatway, NJ, USA, Dec. 1, 2008, pp. 609-616.

Peula J M et al :"Pure Reactive behavior learning using Case Based Reasoning for a vision based 4-legged robot", Robotics and Autonomous Systems, Elsevier Science Publishers, Amsterdam, NL, vol. 57, No. 6-7, Jun. 30, 2009, pp. 688-699.

* cited by examiner

Primary Examiner — Carlos Ortiz Rodriguez
(74) *Attorney, Agent, or Firm* — Mark M. Friedman

(57) **ABSTRACT**

A controller and corresponding method for operating a machine maintains a historical database including at least one sequence of parameters relating to the operating environment, corresponding actions taken, and corresponding outcomes of operation of the machine. The controller searches the database for episodes satisfying relevance criteria relative to a current sequence of parameters and then performs a randomized selection between two or more options, at least one of which is derived from similar episodes from the database. At least one control signal is then output to the machine indicating an action to be performed as determined based on the selected option.

26 Claims, 6 Drawing Sheets

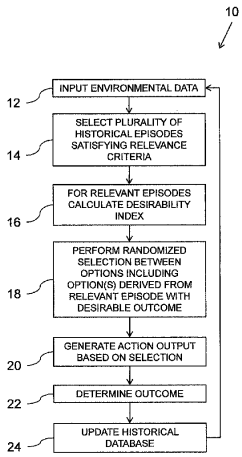


FIG. 1

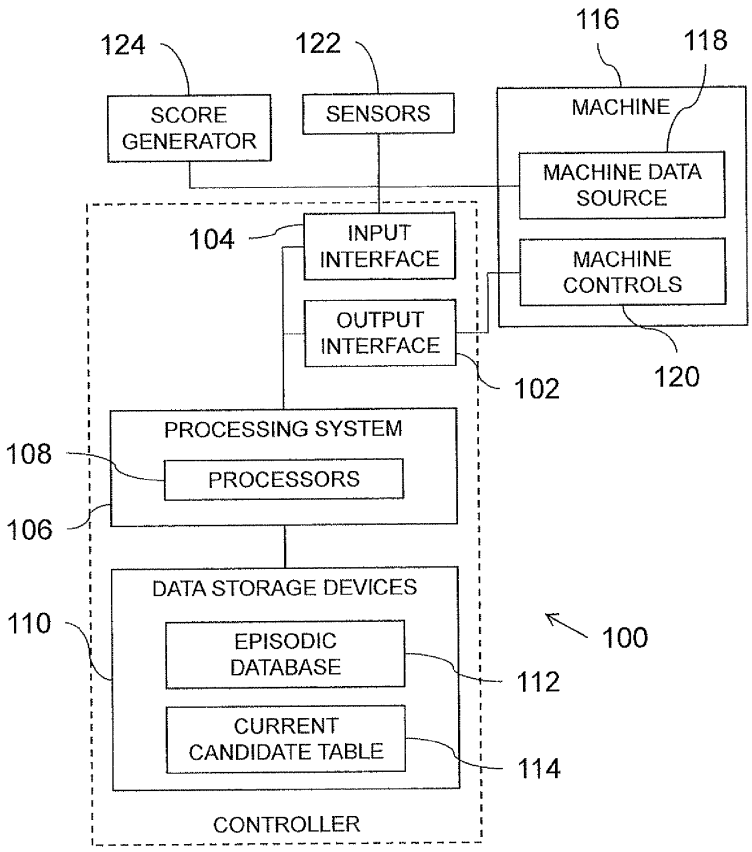

124
122
116
118
SCORE GENERATOR
SENSORS
MACHINE
MACHINE DATA SOURCE
104
INPUT INTERFACE
MACHINE CONTROLS
OUTPUT INTERFACE
102
120
PROCESSING SYSTEM
108
PROCESSORS
106
DATA STORAGE DEVICES
110
100
EPISODIC DATABASE
112
CURRENT CANDIDATE TABLE
114
CONTROLLER

FIG. 2A

112

INDEX	INPUT	ACTION	SCORE
1	A	X	0
2	B	X	1
3	A	Y	2
4	C	Z	0
5	A	Y	1

FIG. 2B

112

INDEX	TYPE	VALUE
1	INPUT	A
2	INPUT	B
3	ACTION	X
4	INPUT	A
5	SCORE	0
6	ACTION	Y
7	INPUT	A
8	SCORE	2

FIG. 3

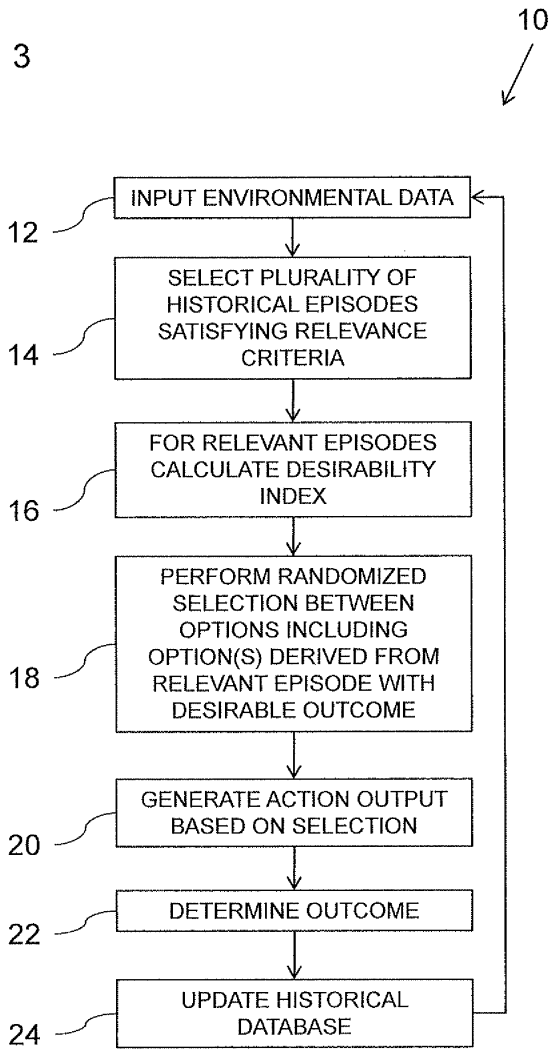

10
INPUT ENVIRONMENTAL DATA
12
SELECT PLURALITY OF HISTORICAL EPISODES SATISFYING RELEVANCE CRITERIA
14
FOR RELEVANT EPISODES CALCULATE DESIRABILITY INDEX
16
PERFORM RANDOMIZED SELECTION BETWEEN OPTIONS INCLUDING OPTION(S) DERIVED FROM RELEVANT EPISODE WITH DESIRABLE OUTCOME
18
GENERATE ACTION OUTPUT BASED ON SELECTION
20
DETERMINE OUTCOME
22
UPDATE HISTORICAL DATABASE
24

FIG. 4

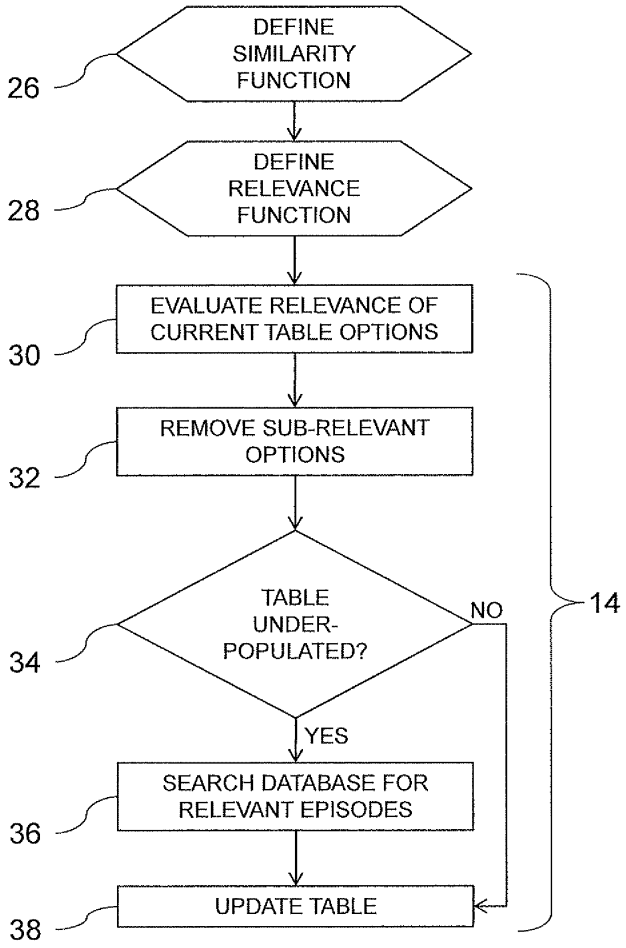

DEFINE SIMILARITY FUNCTION
26
DEFINE RELEVANCE FUNCTION
28
EVALUATE RELEVANCE OF CURRENT TABLE OPTIONS
30
REMOVE SUB-RELEVANT OPTIONS
32
TABLE UNDER-POPULATED?
34
NO
YES
SEARCH DATABASE FOR RELEVANT EPISODES
36
UPDATE TABLE
38
14

FIG. 5

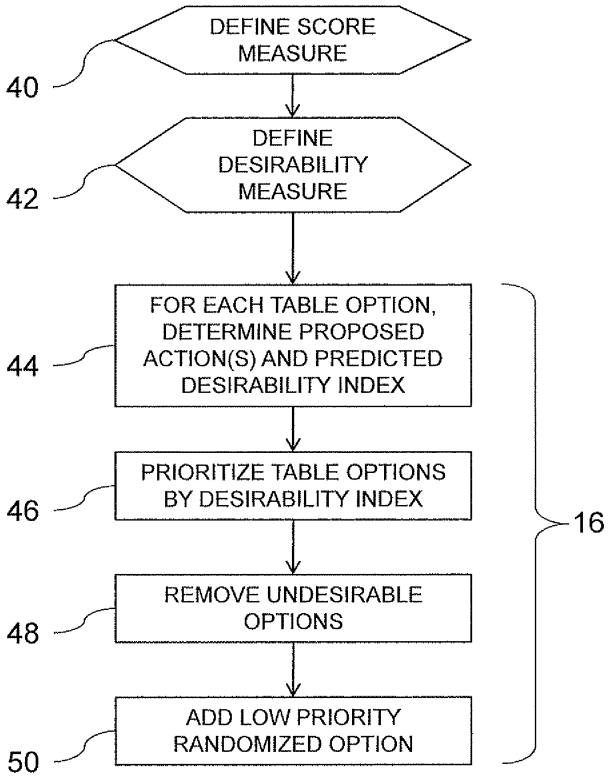
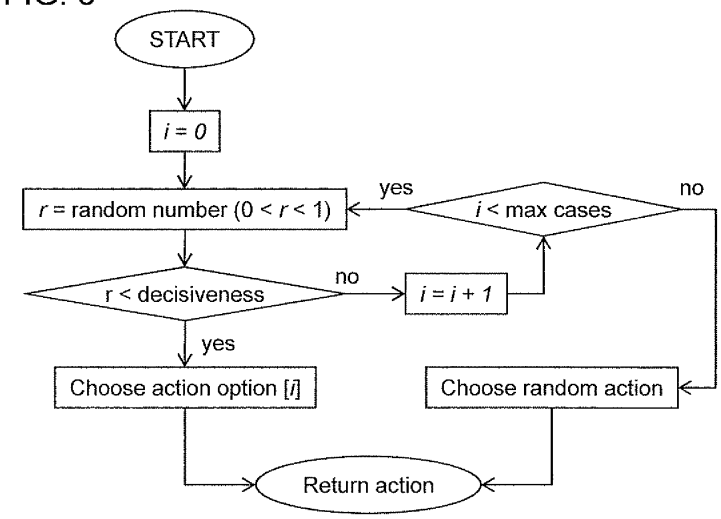

DEFINE SCORE MEASURE
40
DEFINE DESIRABILITY MEASURE
42
FOR EACH TABLE OPTION, DETERMINE PROPOSED ACTION(S) AND PREDICTED DESIRABILITY INDEX
44
PRIORITIZE TABLE OPTIONS BY DESIRABILITY INDEX
46
REMOVE UNDESIRABLE OPTIONS
48
ADD LOW PRIORITY RANDOMIZED OPTION
50
16

FIG. 6


START
i = 0
r = random number (0 < r < 1)
yes
i < max cases
no
r < decisiveness
no
i = i + 1
yes
Choose action option [i]
Choose random action
Return action

CONTROLLER WITH ARTIFICIAL INTELLIGENCE BASED ON SELECTION FROM EPISODIC MEMORY AND CORRESPONDING METHODS

FIELD AND BACKGROUND OF THE INVENTION

The present invention relates to controllers employing artificial intelligence and, in particular, it concerns a controller with artificial intelligence based on selection from stored episodic memory and corresponding methods.

Much emphasis has been put on how to codify "experience" of an artificial intelligence engine into some sort of codified decision-making structure, whether based on neural networks, rules or decision trees. Attempts are made to improve the codified structure based on experience. At any given time, the decision-making structure typically generates a unique output for a given input according to whatever is considered "optimum" for the current state of the decision-making structure.

SUMMARY OF THE INVENTION

The present invention is a controller with artificial intelligence based on selection from stored episodic memory and corresponding methods.

According to an embodiment of the present invention there is provided, a controller for controlling operation of a machine operating in an operating environment, the machine being responsive to control signals to perform a plurality of actions, the controller receiving signals indicative of parameters relating to the operating environment and sufficient to determine an outcome of operation of the machine, the controller comprising: (a) at least one output for providing control signals to the machine to perform selected actions; (b) at least one input for receiving signals indicative of parameters relating to the operating environment; (c) a data storage device containing a historical database including at least one sequence of parameters relating to the operating environment, corresponding actions taken, and corresponding outcomes of operation of the machine; and (d) a processing system including at least one processor, the processing system being in data communication with the data storage device, the at least one output and the at least one input, the processing system being configured to: (i) search within the historical database for similar episodes within the at least one sequence of parameters which satisfy relevance criteria relative to a current sequence of parameters; (ii) perform a randomized selection between a plurality of options, at least one of the options being derived from one of the similar episodes that had a favorable outcome of operation of the machine; and (iii) outputting at least one control signal to the machine via the output to perform a selected action, the control signal being determined as a function of a selected one of the options.

There is also provided, according to an embodiment of the present invention, a method for selecting at least one future action to be performed by a machine, the method comprising the steps of (a) providing a historical database including at least one sequence of parameters relating to the operating environment, corresponding actions taken, and corresponding outcomes of operation of the machine; (b) searching within the historical database for similar episodes within the at least one sequence of parameters which satisfy relevance criteria relative to a current sequence of parameters; (c) performing a randomized selection between a plurality of options, at least one of the options being derived from one of the similar episodes that had a favorable outcome of operation of the machine; and (d) outputting to the machine at least one control signal indicative of at least one action to be performed, the at least one action being determined as a function of a selected one of the options.

According to a further feature of an embodiment of the present invention, the randomized selection is weighted towards selection of a similar episode with a most favorable outcome.

According to a further feature of an embodiment of the present invention, the randomized selection is performed between a plurality of options including a random action option.

According to a further feature of an embodiment of the present invention, at least part of the historical database is provided as a pre-stored database containing a simulated history.

According to a further feature of an embodiment of the present invention, the historical database is updated with recent sequences of parameters relating to the operating environment, corresponding actions taken, and corresponding outcomes of operation of the machine.

According to a further feature of an embodiment of the present invention: (a) a list of episodes is maintained which satisfy the relevance criteria relative to a current sequence of parameters; (b) after receipt of new inputs, the relevance criteria are reevaluated relative to the updated current sequence of parameters and discarding episodes no longer satisfying the relevance criteria; and (c) a search of the historical database is performed intermittently to identify additional episodes of sequences of parameters satisfying the relevance criteria relative to the updated current sequence of parameters for inclusion in the list, wherein the randomized selection is performed using options derived from episodes in the list.

According to a further feature of an embodiment of the present invention, the controlled machine is an additional artificial intelligence controller operating according to a number of operating parameters, and wherein the outputs include a control signal for changing a value of at least one of the operating parameters of the additional artificial intelligence controller.

According to a further feature of an embodiment of the present invention, the searching and the randomized selection are performed according to settings defined by at least one operating parameter, and wherein a value of the at least one operating parameter is changed in response to an input received from an additional artificial intelligence controller.

According to a further feature of an embodiment of the present invention, the parameters relating to the operating environment are sensed, at least in part, by a set of sensors deployed for sensing parameters relating to the operating environment of the machine.

According to a further feature of an embodiment of the present invention, the set of sensors includes a range sensor deployed for sensing a distance from at least part of the machine to an object in the operating environment.

According to a further feature of an embodiment of the present invention, the set of sensors includes an imaging sensor, and wherein the processor system is further configured to perform image processing on imaged sampled by the imaging sensor to derive parameters relating to the operating environment.

According to a further feature of an embodiment of the present invention, the set of sensors includes an audio input,

and wherein the processor system is further configured to perform sound processing on signals sampled by the audio input.

According to a further feature of an embodiment of the present invention, the machine is a virtual machine operating in a computer-generated virtual environment.

There is also provided according to an embodiment of the present invention, a controller for controlling operation of a machine operating in an operating environment, the machine being responsive to control signals to perform a plurality of actions, the controller receiving signals indicative of parameters relating to the operating environment and sufficient to determine an outcome of operation of the machine, the controller comprising: (a) a data storage device; (b) at least one output for providing control signals to the machine to perform selected actions; (c) at least one input for receiving signals indicative of parameters relating to the operating environment; and (d) a processing system including at least one processor, the processing system being in data communication with the data storage device, the at least one output and the at least one input, the processing system being configured to: (i) process a current sequence of parameters together with data from the data storage device to determine a plurality of options satisfying a desirability criterion; (ii) perform a randomized selection between the plurality of options; and (iii) outputting at least one control signal to the machine via the output to perform a selected action, the control signal being determined as a function of a selected one of the options.

There is also provided according to an embodiment of the present invention, a method for selecting at least one future action to be performed by a machine, the method comprising the steps of: (a) processing a current sequence of parameters together with data from the data storage device to determine a plurality of options satisfying a desirability criterion; (b) performing a randomized selection between the plurality of options; and (c) outputting to the machine at least one control signal indicative of at least one action to be performed, the at least one action being determined as a function of a selected one of the options.

BRIEF DESCRIPTION OF THE DRAWINGS

The invention is herein described, by way of example only, with reference to the accompanying drawings, wherein:

FIG. **1** is a block diagram of an embodiment of controller, constructed and operative according to the teachings of the present invention, for controlling a machine;

FIGS. **2**A and **2**B are two tables representing alternative exemplary data structures for a database of episodic history from the controller of FIG. **1**;

FIG. **3** is a high level flow diagram illustrating operation of a controller according to an embodiment of the present invention;

FIG. **4** is a more detailed flow diagram including an expansion of block **14** from the flow diagram of FIG. **3** according to an embodiment of the present invention;

FIG. **5** is a more detailed flow diagram including an expansion of block **16** from the flow diagram of FIG. **3** according to an embodiment of the present invention; and

FIG. **6** is a flow diagram illustrating an implementation of a probabilistic chooser for use in the randomized selection of block **18** of FIG. **3**.

DESCRIPTION OF THE PREFERRED EMBODIMENTS

The present invention is a controller with artificial intelligence based on selection from stored episodic memory and corresponding methods.

The principles and operation of controllers and corresponding methods according to the present invention may be better understood with reference to the drawings and the accompanying description.

Overview of Structure and Operating Principles

Referring now to the drawings, FIG. **1** shows a schematic block diagram representation of a controller, generally designated **100**, constructed and operative according to an embodiment of the present invention, for controlling a machine **116**.

In the embodiment illustrated here, controller **100** includes at least one data storage device **110**, at least one output **102** for providing control signals to machine **116** to perform selected actions, and at least one input **104** for receiving signals indicative of parameters relating to the operating environment of machine **116**. The input signals may originate from a data source **118** within the machine itself, or may be from independent sensors **122** or any other source. The input signals are chosen to be sufficient to determine an outcome of operation of the machine, preferably in the form of one or more "score" parameters. The score may either be derived by controller **100** based on other input parameters, or may be provided from a dedicated external "score generator" **124**, which may optionally include a human teacher.

Controller **100** further includes a processing system **106**, including at least one processor **108**, that is configured to:

- maintain within data storage device **110** a historical database **112** including at least one sequence of parameters relating to the operating environment, corresponding actions taken, and corresponding outcomes of operation of the machine;
- search within the historical database **112** for similar episodes within the sequence of parameters which satisfy relevance criteria relative to a current sequence of parameters;
- perform a randomized selection between a plurality of options, at least one of the options being derived from one of the similar episodes that had a favorable outcome of operation of the machine; and
- output at least one control signal to the machine via the output to perform a selected action, the control signal being determined as a function of a selected one of the options.

Processing system **106** may be implemented using a general purpose computer, modified where necessary to provide appropriate inputs and outputs for the particular application, and operating under control of suitable software that configures the computer to perform the various tasks as defined herein. Code embodied in a computer-readable medium which, when executed on a general purpose computer, is effective to configure the computer to perform the recited steps, also constitutes an aspect of the present invention. The required software may be built around the various processing steps described herein and, being self-explanatory to a person having ordinary skill in the art, will not be described here in detail. Clearly, implementations employing dedicated hardware, or hardware-software combinations referred to as firmware, also fall within the scope of the present invention.

According to one particularly preferred but non-limiting optional implementation, a current set of candidate historical episodes upon which decisions about actions to be taken are based may be stored and managed within a table **114**, to be described in more detail below.

Thus, certain embodiments of the invention may be distinguished from various existing approaches to artificial intelligence in that, rather than codifying a decision-making structure, they maintain an episodic historical record of past

experience, allowing the controller to "remember" what worked in the past in similar situations and to build on that past experience. According to an embodiment of the present invention, a history of all events that happened during the current run, and optionally also in previous runs, is maintained, although it may optionally be manipulated or pruned. This history is used to select the actions to be considered as options for any new situation. No explicit knowledge-base, function, or set or rules is required, although some pre-loaded history may be provided to shorten learning times where appropriate.

Certain embodiments of the invention may additionally, or alternatively, be distinguished from various existing approaches to artificial intelligence in that, rather than consistently selecting a single option considered the "best" according to the current state of the system, they perform a randomized selection which is weighted towards an option identified as the "best" option, but will also sometimes select either another option considered "good", or will experiment with introduction of a randomly (or pseudo-randomly) chosen action. The use of randomized selection renders the controller more human-like in its behavior, and provides adaptability to new situations and changing environments by periodically trying different or new approaches to similar situations.

DEFINITIONS

In this context, "history" or "historical" is used to refer to a record of past events, including inputs, actions taken and scores achieved. The history may be entirely real or may be a simulated history including transplanted "experience" from another controller, and/or a specially manufactured history, to provide a baseline of experience from the start of use. In a simple implementation, the historical database is assumed to be a complete record of the history of the controller, but in practical implementations, it may be pruned in order to reduce data storage requirements and to decrease search time. Pruning can be performed on the basis of discarding the oldest data (particularly suitable for rapidly changing environments), or on the basis of other criteria such as discarding historical episodes which resulted in mediocre desirability of the outcomes, i.e., that are not particularly useful either for re-use or for avoiding past mistakes.

The "events" recorded in the database may be raw input data (such as parameters measured by sensors) or derived data obtained by processing various inputs. Actions taken and outcomes of those actions are preferably also stored as part of the historical database. In some cases, the actions and outcomes are matched up neatly with input data, allowing use of a data structure such as that illustrated in FIG. **2**A. In other cases, the order may be more chaotic, for example, with scores being provided out-of-sync with the actions, and/or without a one-to-one relationship to the actions. In such cases, a data structure such as that of FIG. **2**B is more suitable. In each case, actions and scores appearing during the same period or "episode" as the inputs, or likely to be causally related to the inputs, possibly with a given degree of delay, are referred to as "corresponding" actions and scores.

The term "episode" or "episodic" is used to refer to a time sequence selected from the historical database which can be compared to a current sequence of events to determine its relevance and desirability for the purpose of deciding whether to re-use the sequence for choosing upcoming action(s). An episode is typically identified by a pointer to a location within the historical database which indicates a point in the prior history potentially paralleling the current parameters of the input sequence. A database is referred to as "episodic" if it stores one or more historical sequences rather than just statistical summaries of the data. The database preferably also has a data structure arranged in a manner which facilitates searching and/or retrieval of episodic data, typically from an arbitrary location within the database. In some cases, the historical information may be pruned or otherwise selectively reduced. However, the information is still considered "episodic" so long as there remain a plurality of episodes which are comparable to at least a proportion of the possible input sequences, leaving a decision-making choice to be made by the controller rather than a single-value-output rule.

Reference is made herein to implementations of the present invention as a "controller" for a "machine". The term "machine" here includes a wide range of physical and virtual machines which may be operated under the control of an artificial intelligence engine or "controller". Examples of relevant machines include, but are not limited to, robotic vehicles (submersible, waterborne, land vehicles, airborne and space vehicles), other robotic systems, natural language interaction systems with speech and/or text interfaces, simulators for training people to perform various tasks, and virtual machines or characters operating in virtual environments such as for computer games. A number of specific examples will be described in more detail below.

The "controller" is implemented in various different forms suited to the given application. In cases of robotic devices, the controller may be either implemented as an external controller or may be integrated onto the machine platform. In the case of a virtual machine, the controller is typically implemented as a suitable program module or other virtual object, as is known in the art.

Similarly, the nature of the "operating environment" varies according to the type of application. In a robotic vehicle application, the operating environment typically relates to the physical environment immediately around the machine, and may extend to wide-area information such as geographic position (GPS) and navigation data such as pre-loaded maps. In a speech recognition machine, the environment relates primarily to the sound at the input (microphone) locations. In a virtual-world application, the environment is typically the machine generated environment possibly as modified by other independent computer-operated machines or characters and any human-operated characters.

The term "actions" refers to the entire range of available actions that can be performed by the machine. The available set of actions may vary over time, or as a result of events or previous actions.

The outcomes of operation of the machine is preferably quantified in the form of one or more designated score inputs. The scores reflect an underlying system of evaluation according to which progress towards one or more goals, and/or adherence to one or more principles, can be assessed. Scores may be objectively defined, or may be provided from a subjective source, such as a human observer. Furthermore, a scoring system may vary over time as a situation changes. A score may be a simple number indicating an overall assessment of the outcome, or may have a number of separate components, for example, indicative of progress towards low-level and high-level goals, or short-term and long-term goals. Where different score components are used, they may be represented as a multi-dimensional vector, or may be handled as separate events occurring in the history sequence.

Reference is made to a "current sequence of parameters" which refers to the sequence of parameters or events leading up to the most recent inputs, and possibly including some degree of look-ahead prediction. The current sequence of

parameters is compared against episodes from the historical database to identify similar sequences.

The term "similarity", when used to refer to an individual pair of events, refers to a function which determines a degree of similarity between the events. In certain cases, a simple test of equivalence may be used for discrete variables, or a threshold for continuous variables. However, in many cases, a more complex function may be needed to assess the degree of similarity between a measured parameter in the current sequence and a historical parameter. In some cases, data may be classified as different data types, and different similarity functions may be applied to the different types of data.

The term "relevance" or "relevance criteria" is used to refer to an overall measure of similarity between the current sequence and a historical episode. A basic form of this measure can be a weighted sum of the similarity of pairs of events. In an alternative non-limiting example, for each event, the similarity entered into the calculation may be the similarity of the most similar event in the historical sequence, even if displaced by up to N positions. In the case of N=0, the latter approach simplifies to the former direct comparison of successive pairs of parameters. Differing weights may be assigned for different types of data. In some cases, weights for comparison of actions and scores may be very low or zero, leading to selection of relevant historical episodes based on the operating environment only, while in others, the actions and/or the scores may also play a significant role in the assessment of relevance.

The term "desirability" is used to refer to an overall measure of the expected score over time if following a path of action modeled on a given historical episode, projecting ahead to the future expected scores. The most "desirable" option is thus not necessarily the option which generates the highest short term score, but may instead be the one that will lead to slightly lower scores that are likely to be maintained over time. An option is considered to have a "favorable outcome" if the desirability value satisfies given criteria, such as if the desirability of the given option is one of a number of best-available options from within a group of options being considered, or other criteria defined in comparative or absolute terms. "Most favorable" refers to an option with the highest desirability value from the group of options being considered.

Finally with regard to definitions, reference is made to "randomized selection". In this context, randomized selection refers to any selection which may generate either of at least two results for the same given situation. The randomized selection of embodiments of the present invention is typically performed after a number of options have been sorted by desirability, and is implemented in such a way as to give an increased probability to selection of the most desirable option, and successively decreasing probability to each successively less desirable option. The term "random" is used here loosely to refer to a decision making process which is unpredictable, but does not require that the process is truly random. In certain embodiments, the set of available options in the randomized selection includes a "random action", i.e., randomly selected from a group of available actions not because of any pre-rated desirable result. The random action is preferably allotted a relatively low probability, making it an infrequently chosen path of action. Nevertheless, the occasional use of an unexpected action builds into operation of the controller a degree of adaptability to find new or better solutions to previously encountered situations. Even in the case of a random action, various actions correlated to particularly undesirable outcomes are preferably excluded from the range of options.

Overview of Typical Operation

Turning now to FIGS. **3-5**, these illustrate the overall flow of operation of controller **100** according to a typical implementation, corresponding to a method according to an embodiment of the present invention. Referring first to FIG. **3**, a cycle of operation may be considered to begin with input of some data relating to the operating environment (step **12**) which, together with previously sampled data, generates an updated current sequence of parameters. A plurality of historical episodes satisfying some relevance criteria are selected (step **14**), either directly from the historical database or from a table of candidate options, as will be described in more detail below, and the desirability of choosing a next action according to each of the episodes is calculated (step **16**). Randomized selection is then performed (step **18**) between one or more options derived from relevant episodes with a desirable outcome, optionally together with a random action option. An action output is then generated (step **20**) to actuate the machine to perform the selected action. Score data indicative of the outcome of operation of the machine is then input or determined (step **22**), and the historical database is updated (step **24**). In the "disorderly" data flow scenario of FIG. **2**B, all steps of FIG. **3** occur generally in parallel according to the dynamics of the data flow.

FIG. **4** illustrates in more detail a preferred but non-limiting implementation of selection of relevant historical episodes (step **14**) based on maintaining a table **114** of candidate options. Firstly, as mentioned earlier, assessment of relevance is preferably facilitated by defining one or more similarity functions between individual pairs of data entries (step **26**), and an overall relevance function which combines the individual similarity functions to assess similarity of an sequence of entries (step **28**). The specific exemplary implementation of step **14** then proceeds as follows. At step **30**, the relevance of the current list of options in the table is reevaluated and/or updated based upon the most recently received data, and table entries whose level of relevance has fallen below a certain threshold value are dropped from the table (step **32**). Then, at step **34**, if the removal of less relevant entries from the table has depleted the table below a desired number of entries, a search is performed in historical database **112** to find existing new candidate episodes which have a sufficient level of relevance to the current sequence of parameters to repopulate the table (step **36**) and the table is updated accordingly.

Parenthetically, it should be noted that the sequence illustrated here is appropriate for "normal" operation, but may not always be possible to implement exactly in this form. For example, in the case of a new device starting operation with no previously stored history, the history will initially not contain any matches to the current situation, and random actions will be taken until the device has accumulated sufficient relevant "experience" in the historical database.

It will be noted that the management of candidate entries in a table in this manner greatly reduces the search burden of implementing the present invention, reducing the full database search to an intermittent operation required only when the selection table has become too sparsely populated.

Turning now to FIG. **5**, this illustrates in more detail a preferred but non-limiting implementation of sorting candidate options according to desirability (step **16**), again employing the table **114** of candidate options. Here too, as mentioned earlier, assessment of desirability is facilitated by defining one or more score measures (step **40**) and an overall measure of desirability (step **42**). The notion of similarity is central to these algorithms. In the simplest examples, if two

objects are identical they are considered similar, otherwise not. However, the similarity function can also be any of the following:

A linear vector of weights for comparing vectorial inputs,

A neural-network,

Any other system.

These similarity functions may require tuning/training for every particular application. This can be achieved using genetic algorithms, back-propagation of error, another instance of a controller according to the teachings of the present invention, or any other method. The tuning allows the similarity function to give appropriate weight to the various data of varying importance.

Similarly, any or all of the other operating parameters controlling the algorithm may be tuned using whatever manual or automatic techniques are appropriate to the intended application.

The specific non-limiting exemplary implementation of step **16** then proceeds as follows. Firstly, for each table option, the corresponding proposed next action and the predicted desirability index is calculated (step **44**), and the table is sorted according to the desirability index (step **46**). Preferably, any options with undesirable predicted outcomes are removed from the decision making options, and preferably from the entire table (step **48**). According to certain preferred embodiments, as mentioned above, a low priority randomized option is added as a possible action (step **50**) in order to provide for occasional unpredictable or experimental behavior. The process then proceeds with the randomized selection at step **18** of FIG. **3**.

Example

To illustrate operation of the AI controller of the present invention according to an exemplary non-limiting embodiment, we present a simple experiment according to which the AI controller is taught to differentiate between words and non-word (N-W) sequences of three letters (of the letters: 'D', 'G', 'O'). There are four words made from these letters: 'DOG', 'GOD', 'GOO', 'ODD'.

At each turn the AI is given a letter and it should output if the last three letters form a valid word. If it is correct it is given a score of one, else it is given a score of zero.

Recent history is shown in Table 1:

TABLE 1

897			898			899			900		
Input	Action	Score	Input	Action	Score	Input	Action	Score	Input	Action	Score
'O'	WORD	1	'G'	N-W	1	'O'	N-W	1	'O'	WORD	1

At turn 900 the input is 'O', and the table looks as shown below in Table 2. Note that the events shown are (the beginning of) the predicted future, from which the desirability is extracted. The columns in which we would have seen 'G' 'O' are "to the left" (not shown)—the leftmost column in the table is the one which currently corresponds to the input in iteration No. 900.

TABLE 2

Index	Relevance	Input	Action	Score	Input	Action	Score
1	0.4	'D'	WORD	1	'O'	N-W	1
2	1.3	'O'	WORD	1	'G'	N-W	1
3	1.3	'O'	WORD	1	'G'	N-W	1
4	0.4	'D'	WORD	1	'G'	N-W	1
5	0.4	'D'	WORD	1	'G'	N-W	1
6	0.4	'D'	WORD	1	'G'	N-W	1
7	0.96	'O'	WORD	1	'O'	N-W	1
8	1.6	'O'	WORD	1	'O'	N-W	1
9	1.6	'O'	N-W	0	'G'	WORD	0
10	1.37	'O'	N-W	0	'D'	N-W	1
11	1.6	'O'	N-W	0	'G'	N-W	1
12	1.6	'O'	N-W	0	'G'	N-W	1
13	1.3	'O'	WORD	1	'O'	N-W	1
14	1.07	'O'	N-W	1	'D'	WORD	0
15	1	'O'	N-W	1	'O'	WORD	1
16	1.13	'O'	N-W	1	'G'	N-W	0
17	1.06	'O'	N-W	1	'D'	WORD	1
18	0.85	'O'	N-W	1	'D'	WORD	0
19	1.08	'O'	N-W	1	'D'	WORD	0
20	0.56	'O'	N-W	1	'D'	WORD	0

After sorting for relevance and taking the top 8 we are left with a list of relevant options shown in Table 3:

TABLE 3

8	1.6	'O'	WORD	1	'O'	N-W	1
9	1.6	'O'	N-W	0	'G'	WORD	0
11	1.6	'O'	N-W	0	'G'	N-W	1
12	1.6	'O'	N-W	0	'G'	N-W	1
10	1.37	'O'	N-W	0	'D'	N-W	1
2	1.3	'O'	WORD	1	'G'	N-W	1
3	1.3	'O'	WORD	1	'G'	N-W	1
13	1.3	'O'	WORD	1	'O'	N-W	1

After sorting by desirability and taking the top four we have prioritized preferred options as shown in Table 4:

TABLE 4

8	1.6	'O'	WORD	1	'O'	N-W	1
2	1.3	'O'	WORD	1	'G'	N-W	1
3	1.3	'O'	WORD	1	'G'	N-W	1
13	1.3	'O'	WORD	1	'O'	N-W	1

The probabilistic chooser chose the first line, "8", and the chosen action was WORD.

The score for that was 1, because 'GOO' is a word.

The randomized selection can be performed in many ways. By way of one non-limiting example, FIG. **6** illustrates an implementation of a "probabilistic chooser" algorithm which selects items numbered 0, 1, 2 etc. from a table of prioritized options, with a probability defined by a parameter [decisiveness] (e.g., 0.7) of selecting the first option, [decisiveness]^2 of selecting the second option etc. until it has considered a number of options set by parameter [max cases]. When it reaches the limit without having selected an action, a random action is chosen (subject to optional exclusion of any unacceptable options).

Both operating parameters [decisiveness] and [max cases] may vary over time and/or according to any other algorithm. For example, during the early stages of training of the controller, when historical accumulated data is insufficient to

provide a strong basis for decisions, the probabilistic chooser may be biased more towards exploratory random actions by employing relatively low values for either or both of these parameters. When the controller is well trained through extensive experience, relatively higher values for one or both parameters may be more appropriate, giving a stronger bias towards the solutions that have worked well in the past while maintaining a smaller (but typically non-zero) probability of selecting an exploratory random action. In certain applications, particularly where experimental behavior could be damaging or dangerous, the options for exploratory random actions may be subject to limitation rules or entirely excluded.

Optional Features, Refinements and Variants

Panic mode is based on the observation that people won't choose an option that is much worse than the others. So the algorithm won't choose the worst option harvested from the table, based on the (immediate, or averaged) reward signal.

Although the decision making process described above as a multi-stage process in which a table is populated and maintained, sorted according to "relevance" of the options and then a choice is made according to "desirability", it should be noted that this subdivision is not essential. For example, the relevance may also play a role in the choosing process.

As mentioned above, the AI can be seeded by loading a previously saved history into the history table. The seeded history makes the beginning of the history, and new events are appended to it. This enables carrying knowledge forward from one run to another. It also enables us to seed the AI with human created or procedure created "fore-knowledge".

The Algorithm's learning function can also be turned off, so that the "seeded" knowledge becomes the only usable knowledge—this is useful is application where we don't want the machine to "learn new tricks" in the field, as it may be too dangerous. A pre-loaded set of memories would contain all necessary knowledge for the task-range that the machine would likely encounter.

Fading/pruning memories, creating/concentrating concepts: memories that are not used for a long time can be eliminated, not only for saving on memory space, but mainly for run-time economies. Therefore, the frequency in which memories are used may be monitored. The frequently-used memories become in a sense less a log of a specific event but more a prototypical example that is used as a concept or a habit.

The importance of a memory can be based on its frequency of usage (as above) or on the sequence's importance in terms of getting the algorithm "out of trouble" (the improvement achieved over a period of time).

Use of Multiple Controllers—"Federating"

Using more than one engine to deal with a situation allows for different engines to handle different aspects, when these are conceptually separable. Optionally, it allows for one engine to control the others so as to achieve a complex task. Federating can be implemented in a number of different ways.

Federating hierarchically: engines can be arranged in a hierarchy in the following ways:

A subordinated engine can also decide as an action to pass some input to a superior engine, thereby alerting the "planner" if something out-of-the-ordinary happens during the execution of a plan. "Upward" communication can also allow the lower engine to parse or otherwise recognize the details, and pass up its understanding of the details, like in a system for voice-recognition, natural language processing, and finally substantive response.

The control of the superior engines on the subordinated engines can be done by replacing the memories on which the controlled engine works, or by changing some circumstance in the controlled engine's inputs. The "superior engine" may also change any of the operating parameters of its subordinates, as an expression of its experience of what type of parameters would allow the lower engine of serve the overall purpose better.

Federating in time: different engines can act and calculate in different time magnitudes, for example, one controlling second-by-second, one minute-by-minute, one day-by-day, so that complex tasks can be accomplished.

Federating, all at the same level: each engine looks at a different part of a problem, in a distributed manner, or with some form of overall task coordination, whether based on this technology (hierarchy) or not.

Self modification: each instance to itself or one instance modifying the others, hierarchically or not:

1. The algorithm for what is "similar" can vary;
2. The algorithm for "choosing the best" and adding random action can vary;
3. The policy as to what counts as "subsequently" can vary;
4. Any setting or parameter can vary;
5. The algorithm for calculating or using scores and/or desirability can change, effectively changing or refining what counts as "good".
6. The same machine can have multiple algorithms of each of these categories above in it at the same time, and one of the actions available to it can be to switch between them; for instance, the machine may learn by experience that when something red appears nearby it is best to think short-term, while yellow things are better handled long-term.

Mixed "federations": some of the machines is a federation can be of a different technology, like Neural Networks, etc.

Other Variants

If the function to calculate the "desirability" from the score is constant, run-time efficiency may be improved by caching calculated desirability values previously calculated for a given event vis-à-vis similar events in history, thereby avoiding repetitious calculations.

A specific variation to the "relevance" algorithm could be to prefer more recent experiences. That would allow the machine to change its pattern of behavior and adapt to a changing environment. An opposite preference, for the old, would create a more fixed, perhaps neurotic, behavior. This parameter (the slope and shape of the "curve of preference" as a function of time) can also be subjected to the machine's own control.

More or less than one action can be computed in every iteration.

Although the decision making process has been described above in a particularly preferred implementation based upon use of a table containing a subset of available options considered relevant, it should be noted that alternative implementations may work without such a selection process, instead assigning a weight/relevance to many or all possible sequences. This allows all memories to play some part in the decision making process.

Example

Land Vehicle

An AI controller according to an embodiment of the present invention may be implemented for driving a vehicle, whether an airborne vehicle, land vehicle, waterborne craft or

submergible. For the purpose of one non-limiting example, the controller may be part of a driving system for an automated driverless car or robot courier. For such an implementation, the inputs, outputs and scores may be defined as follows:

INPUTS may include and one or more of the following:

- distance to surrounding objects to front, sides, and/or rear of vehicle, determined, for example, by range sensors;
- speed as determined, for example, by a vehicle speedometer;
- compass bearing and/or GPS location as determined, for example, by a digital compass and/or GPS receiver;
- properties of the driving surface, determined, for example, by one or more on-board accelerometer to indicate bumpiness;
- route planning information;
- a rain sensor;
- a sensor for lighting conditions;
- a data source regarding speed limits;
- a camera with image processing, for example, for identifying street signs;
- vehicle warning indicators such as petrol level, engine temperature, oil/water warnings.

Optionally, various of the inputs can be provided by multi-layer processing of video (and/or audio) inputs. For example, a number of federated controllers, one or more of which operate according to the teachings of the present invention, may be used to:

a) perform image segmentation and/or object recognition to identify traffic signs, other vehicles;

b) perform simultaneous localization and mapping (SLAM) processing to derive vehicle motion and a 3D model of environment;

c) employ the 3D model for short-range navigation to avoid collisions; and

d) employ the 3D model together with GPS and/or compass data for correlation against a map for navigation to the target location.

No one type of input is crucial. For example, use of a video camera plus SLAM may replace distance sensors, or vice versa. In some applications, use of a LIDAR sensor may be advantageous.

For learning basic safe-driving practices, the AI controller can be trained in a virtual environment, or in a controlled real environment, or by transplantation of a suitable basic memory. Where a number of federated controllers are used, the prior training may be provided for all, or only a subset, of the controllers. For certain critical tasks, a "no more learning" option may be implemented in which that controller does not store any new info, and does not take random actions.

ACTIONS: these typically include:

- increase/decrease acceleration;
- increase/decrease braking;
- incremental steering to right/left;
- actuate/deactivate signals right/left;
- controlling steerable video sensors.

Where multiple federated controllers are used, some or all of the controllers may provide an "action" which is an output serving as an input, score or to update an operating parameter of another controller.

SCORE: This typically originates from one or more of two sources:

- "objective" scores come from approaching the destination (closing the distance) and avoiding collisions.
- "subjective" scores can be provided by a human instructor, just as a human being learns to drive.

For handling spatial relationships, polar coordinates are generally the preferred framework for an agent to view its environment. To further simplify processing, a group of objects can be pre-processed into a single entity of a "cluster" which has a characteristic of being, for example, 3, 5, 20-30, etc. members. This cluster is preferably presented to the AI engine as a single entity in the polar space.

Example

Natural Language Interaction

A natural language interaction system is a system which allows people to interact verbally with a computer system as if talking freely to another person. This may be used in a wide range of applications including, but not limited to: a computer input interface; an interactive information system; a customer service system; and a voice operated reservation system.

Typical inputs may include, depending on the level of analysis required, an audio input, such as a microphone or an input arrangement for receiving audio files or signals from a remote audio source (for voice recognition), letters or phonemes for word recognition, and words for the conversational level. Here too, these different levels may be federated between a hierarchy of AI controllers, one or more of which operate according to the teachings of the present invention, each operating on a different level and providing its output as the input to the next level. Clearly, many other inputs may be used according to the particular application. For example, in a customer-service system, additional inputs may be derived from a corporate database, customer records etc.

The actions of the system may be the recognized words, particularly for a computer input system, or may be the responses for a "customer service" type system.

The score is typically provided initially by human trainers, but may also employ "satisfaction indicators" derived from the content and/or tone of the customer's speech.

Example

Computer-Generated Character

The AI controller of the present invention may also be used to operate one or more character in a virtual world, typically for interacting directly with a human user, or with a character within a game that is controlled directly by a human user.

Inputs and outputs for such applications are defined within the framework of the virtual world. Training and scores may be provided by human trainers via suitable inputs, as a separate training process, or possibly during progression of the game.

It will be appreciated that the above descriptions are intended only to serve as examples, and that many other embodiments are possible within the scope of the present invention as defined in the appended claims.

What is claimed is:

1. A controller for controlling operation of a machine operating in an operating environment, the machine being responsive to control signals to perform a plurality of actions, the controller receiving signals indicative of parameters relating to the operating environment and sufficient to determine a score relating to an outcome of operation of the machine, the controller comprising:

(a) at least one output for providing control signals to the machine to perform selected actions;

(b) at least one input for receiving signals indicative of parameters relating to the operating environment;

(c) a data storage device containing a historical database including: at least one sequence of parameters relating to the operating environment; corresponding actions taken; and corresponding scores relating to outcomes of operation of the machine; and

(d) a processing system including at least one processor, said processing system being in data communication with said data storage device, said at least one output and said at least one input, said processing system being configured to:

(i) maintaining a list of episodes, corresponding to sequences identified within the historical database, which satisfy a relevance criteria relative to a current sequence of parameters, said relevance criteria employing a relevance function corresponding to an overall measure of similarity between the current sequence of parameters and the episode derived by combining a plurality of similarity functions, values of said relevance function for each episode in said list being updated based on recently input parameters;

(ii) perform a randomized selection between a plurality of actions or sequences of actions, at least one of said actions or sequences of actions being derived from an episode in said list that had a favorable outcome of operation of the machine; and

(iii) outputting at least one control signal to the machine via said at least one output to perform the selected action or sequence of actions.

2. The controller of claim **1**, wherein said processing system is configured to perform said randomized selection wherein said randomized selection is weighted towards selection of a similar episode with a most favorable outcome.

3. The controller of claim **1**, wherein said processing system is configured to perform said randomized selection between a plurality of actions or sequences of actions including a random action.

4. The controller of claim **1**, wherein at least part of said historical database is provided as a pre-stored database containing a simulated history.

5. The controller of claim **1**, wherein said processing system is further configured to update said historical database with: recent sequences of parameters relating to the operating environment; corresponding actions taken; and corresponding outcomes of operation of the machine.

6. The controller of claim **1**, wherein said processing system is further configured to:

(a) after receipt of new inputs, reevaluate said relevance criteria relative to an updated current sequence of parameters and discard episodes no longer satisfying said relevance criteria; and

(c) intermittently perform a search of said historical database to identify additional episodes satisfying said relevance criteria relative to the updated current sequence of parameters for inclusion in said list.

7. The controller of claim **1**, wherein the controlled machine is an additional artificial intelligence controller operating according to a number of operating parameters, and wherein said outputs include a control signal for changing a value of at least one of said operating parameters of said additional artificial intelligence controller.

8. The controller of claim **1**, wherein said processing system performs said maintaining a list and said randomized selection according to settings defined by at least one operating parameter, and wherein said processing system is responsive to an input received from an additional artificial intelligence controller to change a value of said at least one operating parameter.

9. The controller of claim **1**, further comprising a set of sensors associated with said at least one input and deployed for sensing parameters relating to the operating environment of the machine.

10. The controller of claim **9**, wherein said set of sensors includes a range sensor deployed for sensing a distance from at least part of the machine to an object in the operating environment.

11. The controller of claim **9**, wherein said set of sensors includes an imaging sensor, and wherein said processing system is further configured to perform image processing on images sampled by said imaging sensor to derive parameters relating to the operating environment.

12. The controller of claim **9**, wherein said set of sensors includes an audio input, and wherein said processor system is further configured to perform sound processing on signals sampled by said audio input.

13. The controller of claim **1**, wherein the machine is a virtual machine operating in a computer-generated virtual environment.

14. A method for selecting at least one future action to be performed by a machine, the method receiving as an input signals indicative of parameters relating to an operating environment of the machine and sufficient to determine a score relating to an outcome of operation of the machine, the method comprising the steps of

(a) providing a historical database including: at least one sequence of parameters relating to the operating environment; corresponding actions taken; and corresponding scores relating to outcomes of operation of the machine;

(b) maintaining a list of episodes, corresponding to sequences identified within the historical database, which satisfy a relevance criteria relative to a current sequence of parameters, said relevance criteria employing a relevance function corresponding to an overall measure of similarity between the current sequence of parameters and the episode derived by combining a plurality of similarity functions, values of said relevance function for each episode in said list being updated based on recently input parameters;

(c) performing a randomized selection between a plurality of actions or sequences of actions, at least one of said actions or sequences of actions being derived from an episode in said list that had a favorable outcome of operation of the machine; and

(d) outputting to the machine at least one control signal indicative of the selected action or sequence of actions to be performed.

15. The method of claim **14**, wherein said randomized selection is weighted towards selection of a similar episode with a most favorable outcome.

16. The method of claim **14**, wherein said randomized selection is performed between a plurality of actions or sequences of actions including a random action.

17. The method of claim **14**, wherein at least part of said historical database is provided as a pre-stored database containing a simulated history.

18. The method of claim **14**, further comprising updating said historical database with: recent sequences of parameters relating to the operating environment; corresponding actions taken; and corresponding outcomes of operation of the machine.

19. The method of claim **14**, wherein said maintaining a list of episodes comprises:

(a) after receipt of new inputs, reevaluating said relevance criteria relative to an updated current sequence of parameters and discarding episodes no longer satisfying said relevance criteria; and

(c) intermittently performing a search of said historical database to identify additional episodes satisfying said relevance criteria relative to the updated current sequence of parameters for inclusion in said list.

20. The method of claim **14**, wherein the machine is an additional artificial intelligence controller operating according to a number of operating parameters, and wherein said at least one control signal include a control signal for changing a value of at least one of said operating parameters of said additional artificial intelligence controller.

21. The method of claim **14**, wherein said maintaining a list and said randomized selection are performed according to settings defined by at least one operating parameter, and wherein a value of said at least one operating parameter is changed in response to an input received from an additional artificial intelligence controller.

22. The method of claim **14**, wherein said parameters relating to the operating environment are sensed, at least in part, by a set of sensors deployed for sensing parameters relating to the operating environment of the machine.

23. The method of claim **22**, wherein said set of sensors includes a range sensor deployed for sensing a distance from at least part of the machine to an object in the operating environment.

24. The method of claim **22**, wherein said set of sensors includes an imaging sensor, and wherein said processing system is further configured to perform image processing on images sampled by said imaging sensor to derive parameters relating to the operating environment.

25. The method of claim **22**, wherein said set of sensors includes an audio input, and wherein said processing system is further configured to perform sound processing on signals sampled by said audio input.

26. The method of claim **14**, wherein the machine is a virtual machine operating in a computer-generated virtual environment.

* * * * *

10 Bibliography

Agre, P. E. (1997). Toward a Critical Technical Practice: Lessons Learned in Trying to Reform AI. In G. Bowker, S. L. Star, L. Gasser, & W. Turner (Eds.), *Social Science, Technical Systems, and Cooperative Work: Beyond the Great Divide* (pp. 131–157). Psychology Press.

Altmann, G. T. M., & Dienes, Z. (1999). Rule Learning by Seven-Month-Old Infants and Neural Networks. *Science*, *284*(5416), 875–875. https://doi.org/10.1126/science.284.5416.875a

Amazon Prime Air. (2016). Retrieved September 24, 2015, from http://www.amazon.com/b?node=8037720011

Ariely, D. (2009). *Predictably irrational: the hidden forces that shape our decisions*. London: Harper.

Aristotle. (2009). *Nicomachean Ethics*. (W. D. Ross, Trans.). Retrieved from http://classics.mit.edu/Aristotle/nicomachaen.1.i.html

Augier, M., & March, J. G. (2001). Remembering Herbert A. Simon (1916-2001). *Public Administration Review*, *61*(4), 396–402.

Austin, J. L., & Warnock, G. J. (1964). *Sense and Sensibilia*. Oxford University Press.

Baddeley, B., Graham, P., Husbands, P., & Philippides, A. (2012). A Model of Ant Route Navigation Driven by Scene Familiarity. *PLoS Comput Biol*, *8*(1), e1002336. https://doi.org/10.1371/journal.pcbi.1002336

Bannister, R. C. (1991). *Sociology and Scientism: The American Quest for Objectivity, 1880-1940*. Chapel Hill: The University of North Carolina Press.

Bar-Hillel, Y. (2003). The Present Status of Automatic Translation of Languages. In S. Nirenburg & H. L. Somers (Eds.), *Readings in Machine Translation* (pp. 45–77). MIT Press.

Beyer, C. (2015). Edmund Husserl. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of*

Philosophy (Summer 2015). Retrieved from http://plato.stanford.edu/archives/sum2015/entries/husserl/

Bloch, M. L. B. (1953). *The historian's craft*. New York, N.Y: Vintage Books. Retrieved from http://capitadiscovery.co.uk/sussex-ac/items/1081954

Boden, M. (2008). *Mind as Machine: A History of Cognitive Science*. OUP Oxford.

Bogen, J. (2014). Theory and Observation in Science. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2014). Retrieved from http://plato.stanford.edu/archives/sum2014/entries/science-theory-observation/

Bolter, J. D. (1984). *Turing's man: western culture in the computer age*. London: Duckworth. Retrieved from http://capitadiscovery.co.uk/sussex-ac/items/216579

Borges, J. L. (2001). The Total Library. In E. Allen & S. Levine (Trans.), *The Total Library: Non-Fiction 1922-1986* (New Ed edition, pp. 214–216). Penguin Classics.

Bourdeau, M. (2014). Auguste Comte. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2014). Retrieved from http://plato.stanford.edu/archives/win2014/entries/comte/

Bower, J. M., & Bolouri, H. (2001). *Computational Modeling of Genetic and Biochemical Networks*. MIT Press.

Breuer, Y., & Shavit, E. (2014). *Hilarious Hebrew: The Fun and Fast Way to Learn the Language*. Pitango Publishing.

Bringsjord, S. (2008). The Logicist Manifesto. Retrieved from http://kryten.mm.rpi.edu/SB_LAI_Manifesto_091808.pdf

Broekens, J., Heerink, M., & Rosendal, H. (2009). Assistive social robots in elderly care: a review. *Gerontechnology*, *8*(2). https://doi.org/10.4017/gt.2009.08.02.002.00

Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence*, *47*(1–

3), 139–159. https://doi.org/10.1016/0004-3702(91)90053-M

Brooks, R. A., Breazeal, C., Marjanović, M., Scassellati, B., & Williamson, M. M. (1999). The Cog Project: Building a Humanoid Robot. In C. L. Nehaniv (Ed.), *Computation for Metaphors, Analogy, and Agents* (pp. 52–87). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-48834-0_5

Brown, A. (2010). *The rise and fall of Communism*. London: Vintage.

Byers, E. S., Purdon, C., & Clark, D. A. (1998). Sexual intrusive thoughts of college students. *Journal of Sex Research*, *35*(4), 359–369. https://doi.org/10.1080/00224499809551954

Byrne, A. (2005). Introspection. *Philosophical Topics*, *33*(1), 79–104.

Carr, J. E. (1985). Ethno-Behaviorism and the Culture-Bound Syndromes: the Case of Amok. In R. C. Simons & C. C. Hughes (Eds.), *The Culture-Bound Syndromes* (pp. 199–223). Springer Netherlands. Retrieved from http://link.springer.com/chapter/10.1007/978-94-009-5251-5_20

Carroll, L. (1895). What the Tortoise said to Achilles. *Mind*, *4*(14), 278–280.

Chomsky, N. (1959). A review of BF Skinner's Verbal Behavior. *Language*, *35*(1), 26–58.

Chrisley, R. (2003). Embodied artificial intelligence. *Artificial Intelligence*, *149*(1), 131–150. https://doi.org/10.1016/S0004-3702(03)00055-9

Clark, A. (1998). *Being There: Putting Brain, Body, and World Together Again*. MIT Press.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Clark, A., & Chalmers, D. (1998). The Extended Mind. *Analysis*, *58*(1), 7–19.

Cole, R. A. (1973). Listening for mispronunciations: A measure of what we hear during

speech. *Perception & Psychophysics*, *13*(1), 153–156. https://doi.org/10.3758/BF03207252

Collins, S. H., & Ruina, A. (2005). A Bipedal Walking Robot with Efficient and Human-Like Gait. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation, 2005. ICRA 2005* (pp. 1983–1988). https://doi.org/10.1109/ROBOT.2005.1570404

Costall, A. (2004). From Darwin to Watson (and Cognitivism) and Back Again: The Principle of Animal-Environment Mutuality. *Behavior and Philosophy*, *32*(1), 179–195.

Costall, A. (2006). “Introspectionism” and the mythical origins of scientific psychology. *Consciousness and Cognition*, *15*(4), 634–654. https://doi.org/10.1016/j.concog.2006.09.008

Cowie, F. (2010). Innateness and Language. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2010). Retrieved from http://plato.stanford.edu/archives/sum2010/entries/innateness-language/

Dennett, D. (2003). Who’s On First? Heterophenomenology Explained. *Journal of Consciousness Studies*, *10*(9–10), 19–30.

Dennett, D. C. (1989). *The Intentional Stance*. MIT Press.

Deryugina, O. V. (2010). Chatterbots. *Scientific and Technical Information Processing*, *37*(2), 143–147. https://doi.org/10.3103/S0147688210020097

Descartes. (1952). *The Meditations and selections from The Principles*. (Veitch, John, Trans.). La Salle, Illinois: Open Court.

Dowe, D. L. (2013). Introduction to Ray Solomonoff 85th Memorial Conference. In D. L. Dowe (Ed.), *Algorithmic Probability and Friends. Bayesian Prediction and Artificial Intelligence* (pp. 1–36). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-44958-1_1

Dreyfus, H. L. (1979). *What computers can't do / The limits of artificial intelligence* (Revised). New York: Harper & Row.

Dreyfus, H. L. (1996). Response to my critics. *Artificial Intelligence*, *80*(1), 171–191. https://doi.org/10.1016/0004-3702(95)00088-7

Dreyfus, H. L. (2007). Why Heideggerian AI failed and how fixing it would require making it more Heideggerian. *Artificial Intelligence*, *171*(18), 1137–1160. https://doi.org/10.1016/j.artint.2007.10.012

Dreyfus, H. L. (2012). A history of first step fallacies. *Minds and Machines*, *22*(2), 87–99.

Dreyfus, H. L., & Dreyfus, S. E. (1986). *Mind over machine: the power of human intuition and expertise in the era of the computer*. New York: Free Press.

Edwards, P. N. (1997). *The Closed World: Computers and the Politics of Discourse in Cold War America*. MIT Press.

Ericsson, K. A., & Simon, H. A. (1981). Sources of evidence on cognition: An historical overview. In Merluzzi, Glass, & Genest (Eds.), *Cognitive Assesment*. Carnegie-Mellon University, Department of Psychology. Retrieved from http://octopus.library.cmu.edu/cgi-bin/tiff2pdf/simon/box00067/fld05162/bdl0001/doc0001/simon.pdf

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: verbal reports as data* (2nd ed.). Cambridge, Mass: MIT P. Retrieved from http://capitadiscovery.co.uk/sussex-ac/items/543010

Fantl, J. (2014). Knowledge How. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2014). Retrieved from http://plato.stanford.edu/archives/fall2014/entries/knowledge-how/

Feigenbaum, E. A. (1989). What hath Simon wrought. *Complex Information Processing: The Impact of Herbert A. Simon*, 165–182.

Feynman, R. (1988). Richard Feynman's blackboard at time of his death | Caltech. Retrieved October 21, 2015, from http://caltech.discoverygarden.ca/islandora/object/ct1%3A483

Franssen, M., Lokhorst, G.-J., & van de Poel, I. (2013). Philosophy of Technology. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2013). Retrieved from http://plato.stanford.edu/archives/win2013/entriesechnology/

Freed, S. (2013). Practical Introspection as Inspiration for AI. In V. C. Müller (Ed.), *Philosophy and Theory of Artificial Intelligence* (pp. 167–177). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-31674-6_12

Froese, T. (2011). Validating and Calibrating First- and Second-person Methods in the Science of Consciousness. *Journal of Consciousness Studies*, *18*(2), 38–64.

Fromm, E. (2011). *Escape from Freedom*. Place of publication not identified: Ishi Press.

Gadamer, H.-G. (1976). *Philosophical hermeneutics*. (D. E. Linge, Trans.). Berkeley, Calif: U. of California P. Retrieved from http://capitadiscovery.co.uk/sussex-ac/items/37515

Gadamer, H.-G. (1979). *Truth and method* (2nd ed.). London: Sheed and Ward. Retrieved from http://capitadiscovery.co.uk/sussex-ac/items/40876

Gadamer, H.-G. (2004). *Truth and method* (2nd, ed ed.). London ; New York: Continuum.

Gallagher, S., & Zahavi, D. (2012). *The phenomenological mind*. London; New York: Routledge.

Gamez, D. (2008). Progress in machine consciousness. *Consciousness and Cognition*, *17*(3), 887–910. https://doi.org/10.1016/j.concog.2007.04.005

Goffman, E. (1971). *The presentation of self in everyday life*. Harmondsworth: Penguin.

Goldie, P. (2012). *The Mess Inside: Narrative, Emotion, and the Mind*. Oxford: OUP

Oxford.

Heidegger, M. (1962). *Being and time*. (J. Macquarrie & E. Robinson, Trans.). Malden, MA; Oxford: Blackwell.

Heidegger, M. (2009). The Question Concerning Technology. In C. Hanks (Ed.), *Technology and Values: Essential Readings* (pp. 99–113). John Wiley & Sons.

Hesslow, G. (2012). The current status of the simulation theory of cognition. *Brain Research*, *1428*, 71–79. https://doi.org/10.1016/j.brainres.2011.06.026

Hockings, N., Iravani, P., & Bowen, C. R. (2014). Artificial ligamentous joints: Methods, materials and characteristics. In *Humanoids* (pp. 20–26). Retrieved from http://people.bath.ac.uk/nch28/pdfs/Artificial%20Ligamentous%20Joints%20-%20Methods%20Materials%20and%20Characteristics.pdf

Hurlburt, R. T. (2011). *Investigating Pristine Inner Experience: Moments of Truth*. New York: Cambridge University Press.

Hurlburt, R. T., Heavey, C. L., & Kelsey, J. M. (2013). Toward a phenomenology of inner speaking. *Consciousness and Cognition*, *22*(4), 1477–1494. https://doi.org/10.1016/j.concog.2013.10.003

Hyslop, A. (2014). Other Minds. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2014). Retrieved from http://plato.stanford.edu/archives/spr2014/entries/other-minds/

IBM. (2014, March 20). IBM Collaboration Solutions software - Lotus software - United Kingdom [CT503]. Retrieved October 10, 2014, from http://www-01.ibm.com/software/uk/lotus/

Isensee, P. (2001). Genuine Random Number Generation. *Game Programming Gems 2*, 127.

Ismael, J. (2015). Quantum Mechanics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2015). Retrieved from

http://plato.stanford.edu/archives/spr2015/entries/qm/

Jack, A., & Roepstorff, A. (Eds.). (2003). *Trusting the Subject: v. 1*. Exeter etc.: Imprint Academic.

Jack, A., & Roepstorff, A. (Eds.). (2004). *Trusting the Subject: v. 2*. Exeter, UK: Imprint Academic.

Johansson, P., Hall, L., Sikström, S., Tärning, B., & Lind, A. (2006). How something can be said about telling more than we can know: On choice blindness and introspection. *Consciousness and Cognition*, *15*(4), 673–692.

Klotzko, A. J. (2001). *The Cloning Sourcebook*. Oxford University Press, USA.

Knapp, S. (2008). Artificial Intelligence: Past, Present, and Future. Retrieved December 9, 2015, from http://www.dartmouth.edu/~vox/0607/0724/ai50.html

Laird, J. E., & Rosenbloom, P. (1996). The evolution of the Soar cognitive architecture. *Mind Matters: A Tribute to Allen Newell*, 1–50.

Langley, P. (2006). *Intelligent behavior in humans and machines*. Technical Report). Computational Learning Laboratory, CSLI, Stanford University. Retrieved from http://lyonesse.stanford.edu/~langley/papers/ai50.dart.pdf

Lenat, D. B., Prakash, M., & Shepherd, M. (1985). CYC: Using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks. *AI Magazine*, *6*(4), 65.

Malpas, J. (2013). Hans-Georg Gadamer. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2013). Retrieved from http://plato.stanford.edu/archives/sum2013/entries/gadamer/

Mandik, P. (2001). Mental representation and the subjectivity of consciousness. *Philosophical Psychology*, *14*(2), 179–202. https://doi.org/10.1080/09515080120051553

Markram, H. (2006). The Blue Brain Project. *Nature Reviews Neuroscience*, *7*(2), 153–

160. https://doi.org/10.1038/nrn1848

Markram, H. (2012). The Human Brain Project. *Scientific American*, *306*(6), 50–55. https://doi.org/10.1038/scientificamerican0612-50

Matthews, M. R. (1994). *Science Teaching: The Role of History and Philosophy of Science*. Psychology Press.

McCorduck, P. (2004). *Machines who think: a personal inquiry into the history and prospects of artificial intelligence* (25th anniversary update). Natick, Mass: A.K. Peters.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115–133. https://doi.org/10.1007/BF02478259

McHugh, J., & Minsky, M. (2003, August 1). Why A.I. Is Brain-Dead. Retrieved January 8, 2016, from http://www.wired.com/2003/08/why-a-i-is-brain-dead/

McLeod, P., Reed, N., & Dienes, Z. (2003). Psychophysics: How fielders arrive in time to catch the ball. *Nature*, *426*(6964), 244–245.

McNeill, D., & Freiberger, P. (1994). *Fuzzy Logic: The Revolutionary Computer Technology that Is Changing Our World* (1st edition). New York: Touchstone / Simon & Schuster.

Mill, J. S. (2013). *Auguste Comte and Positivism*. CreateSpace Independent Publishing Platform.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97. https://doi.org/10.1037/h0043158

Minsky, M. (1991). Logical Versus Analogical or Symbolic Versus Connectionist or Neat Versus Scruffy. *AI Mag.*, *12*(2), 34–51.

Mladenić, D., & Bradeško, L. (2012). A survey of chatbot systems through a Loebner

prize competition [Conference or Workshop Item]. Retrieved July 17, 2014, from http://eprints.pascal-network.org/archive/00009729/

Mould, R. F. (1998). The discovery of radium in 1898 by Maria Sklodowska-Curie (1867-1934) and Pierre Curie (1859-1906) with commentary on their life and times. *The British Journal of Radiology*, *71*(852), 1229–1254. https://doi.org/10.1259/bjr.71.852.10318996

Müller, V. C. (2009). Pancomputationalism: Theory or metaphor? *The Relevance of Philosophy for Information Science. Berlin: Springer P. Forthcoming*. Retrieved from http://www.typos.de/pdf/2008_Paderborn_Pancomputationalism.pdf

Nagel, T. (1974). What Is It Like to Be a Bat? *The Philosophical Review*, *83*(4), 435–450. https://doi.org/10.2307/2183914

Neisser, U. (1967). *Cognitive psychology*. New York: Appleton-Century-Crofts. Retrieved from http://capitadiscovery.co.uk/sussex-ac/items/27273

Newell, A., & Simon, H. A. (1956). The logic theory machine–A complex information processing system. *IRE Transactions on Information Theory*, *2*(3), 61–79. https://doi.org/10.1109/TIT.1956.1056797

Newell, A., & Simon, H. A. (1961a). Computer simulation of human thinking. *Science*. Retrieved from http://psycnet.apa.org/?fa=main.doiLanding&uid=1962-05907-001.

Newell, A., & Simon, H. A. (1961b). *GPS, a program that simulates human thought*. Defense Technical Information Center. Retrieved from http://octopus.library.cmu.edu/cgi-bin/tiff2pdf/simon/box00064/fld04907/bdl0001/doc0001/simon.pdf

Nietzsche. (1889). Full text of “The Will to Power.” Retrieved October 30, 2014, from https://archive.org/stream/TheWillToPower-Nietzsche/will_to_power-nietzsche_djvu.txt

Nilsson, N. J. (2010). The Quest for Artificial Intelligence. Retrieved March 20, 2015, from http://www.cambridge.org/gb/academic/subjects/computer-science/artificial-intelligence-and-natural-language-processing/quest-artificial-intelligence

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231–259. https://doi.org/10.1037/0033-295X.84.3.231

Nobelprize.org. (1978). The Prize in Economics 1978 - Press Release. Retrieved from http://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/1978/press.html

O'Regan, J. K. (2011). *Why Red Doesn't Sound Like a Bell: Understanding the feel of consciousness*. Oxford University Press.

Overgaard, M. (2006). Introspection in Science. *Consciousness and Cognition*, *15*(4), 629–633. https://doi.org/10.1016/j.concog.2006.10.004

Overgaard, M. (2008). Introspection. *Scholarpedia*, *3*(5), 4953. https://doi.org/10.4249/scholarpedia.4953

Payne, S. J., & Squibb, H. R. (1990). Algebra mal-rules and cognitive accounts of error. *Cognitive Science*, *14*(3), 445–481. https://doi.org/10.1016/0364-0213(90)90019-S

Pear, J. J. (2007). *A Historical and Contemporary Look at Psychological Systems* (1 edition). Mahwah, N.J: Psychology Press.

Piccinini, G. (2004). The First Computational Theory of Mind and Brain: A Close Look at Mcculloch and Pitts's "Logical Calculus of Ideas Immanent in Nervous Activity." *Synthese*, *141*(2), 175–215. https://doi.org/10.1023/B:SYNT.0000043018.52445.3e

Quine, W. van O. (1976). Two Dogmas of Empiricism. In S. G. Harding (Ed.), *Can*

Theories be Refuted? (pp. 41–64). Springer Netherlands. Retrieved from http://link.springer.com/chapter/10.1007/978-94-010-1863-0_2

Raibert, M., Blankespoor, K., Nelson, G., Playter, R., & others. (2008). Bigdog, the rough-terrain quadruped robot. In *Proceedings of the 17th World Congress* (pp. 10823–10825). Retrieved from http://web.unair.ac.id/admin/file/f_7773_bigdog.pdf

Ramberg, B., & Gjesdal, K. (2014). Hermeneutics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2014). Retrieved from http://plato.stanford.edu/archives/win2014/entries/hermeneutics/

Ravenscroft, I. (2010). Folk Psychology as a Theory. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2010). Retrieved from http://plato.stanford.edu/archives/fall2010/entries/folkpsych-theory/

Rayner, K., White, S. J., Johnson, R. L., & Liversedge, S. P. (2006). Raeding Wrods With Jubmled Lettres There Is a Cost. *Psychological Science*, *17*(3), 192–193. https://doi.org/10.1111/j.1467-9280.2006.01684.x

Resnick, M. (1993). Behavior Construction Kits. *Commun. ACM*, *36*(7), 64–71. https://doi.org/10.1145/159544.159593

Reutlinger, A., Schurz, G., & Hūttemann, A. (2014). Ceteris Paribus Laws. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2014). Retrieved from http://plato.stanford.edu/archives/spr2014/entries/ceteris-paribus/

Richtel, M., & Dougherty, C. (2015, September 1). Google's Driverless Cars Run Into Problem: Cars With Drivers. *The New York Times*. Retrieved from http://www.nytimes.com/2015/09/02/technology/personaltech/google-says-its-not-the-driverless-cars-fault-its-other-drivers.html

Robertson, J. (2007). Robo Sapiens Japanicus: Humanoid Robots and the Posthuman Family. *Critical Asian Studies*, *39*(3), 369–398.

https://doi.org/10.1080/14672710701527378

Romano, C. (2009, October 18). Heil Heidegger! *The Chronicle of Higher Education*. Retrieved from http://chronicle.com/article/Heil-Heidegger-/48806/

Rothenberg, A. (1995). Creative Cognitive Processes in Kekulé's Discovery of the Structure of the Benzene Molecule. *The American Journal of Psychology*, *108*(3), 419–438. https://doi.org/10.2307/1422898

Russell, B. (1952). Is there a God? *Why I Am Not a Christian*.

Russell, S., & Norvig, P. (2013). *Artificial Intelligence: A Modern Approach* (3 edition). Harlow: Pearson.

Russell, S., & Norvig, P. (2016). 1293 Schools Worldwide That Have Adopted AIMA. Retrieved January 10, 2016, from http://aima.cs.berkeley.edu/adoptions.html

Safonov, Y. G., & Prokof'ev, V. Y. (2006). Gold-bearing reefs of the Witwatersrand Basin: A model of synsedimentation hydrothermal formation. *Geology of Ore Deposits*, *48*(6), 415–447. https://doi.org/10.1134/S1075701506060018

Schank, R. C. (1982). *Dynamic Memory: A Theory of Learning in Computers and People*. NY: Cambridge.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals, and understanding: an enquiry into human knowledge structures*. Erlbaum. Retrieved from http://capitadiscovery.co.uk/sussex-ac/items/38886

Schickore, J. (2014). Scientific Discovery. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2014). Retrieved from http://plato.stanford.edu/archives/spr2014/entries/scientific-discovery/

Schwitzgebel, E. (2004). Introspective Training Apprehensively Defended: Reflections on Titchener's Lab Manual. *Journal of Consciousness Studies*, *11*(7–8), 58–76.

Schwitzgebel, E. (2012). Introspection. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2012). Retrieved from

http://plato.stanford.edu/archives/win2012/entries/introspection/

Schwitzgebel, E. (2014). Introspection. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2014). Retrieved from http://plato.stanford.edu/archives/sum2014/entries/introspection/

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–424. https://doi.org/10.1017/S0140525X00005756

Searle, J. R. (1992). *The Rediscovery of the Mind* (Massachusetts Institute of Technology edition). Cambridge, Mass: A Bradford Book.

Seth, A. K. (2010). The Grand Challenge of Consciousness. *Frontiers in Psychology*, *1*. https://doi.org/10.3389/fpsyg.2010.00005

Shanon, B. (2008). *Representational and the Presentational: An Essay on Cognition and the Study of Mind* (2nd edition). Exeter, UK ; Charlottesville, VA: Imprint Academic.

Sharkey, A., & Sharkey, N. (2011). Children, the Elderly, and Interactive Robots. *IEEE Robotics Automation Magazine*, *18*(1), 32–38. https://doi.org/10.1109/MRA.2010.940151

Simon, H. A. (1955). A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, *69*(1), 99–118. https://doi.org/10.2307/1884852

Simon, H. A. (1976). *Administrative behavior: a study of decision-making processes in administrative organization* (3rd ed.). London: Collier Macmillan. Retrieved from http://capitadiscovery.co.uk/sussex-ac/items/38710

Simon, H. A. (1981). *The sciences of the artificial* (2nd ed..). Cambridge, Mass: MITP.

Simon, H. A. (1989). The scientist as problem solver. *Complex Information Processing: The Impact of Herbert A. Simon*, 375–398.

Simon, H. A. (1996a). *Models of my life*. Cambridge, Mass: MIT P. Retrieved from http://capitadiscovery.co.uk/sussex-ac/items/547214

Simon, H. A. (1996b). *The sciences of the artificial* (3rd ed.). Cambridge, Mass: MIT Press. Retrieved from http://capitadiscovery.co.uk/sussex-ac/items/546838

Simon, H. A., & Newell, A. (1958). Heuristic Problem Solving: The Next Advance in Operations Research. *Operations Research*, *6*(1), 1–10.

Skinner, B. F. (1987). Whatever happened to psychology as the science of behavior? *American Psychologist*, *42*(8), 780–786. https://doi.org/10.1037/0003-066X.42.8.780

Smith, B. C. (2005, January 31). Digital Future: Meaning of Digital. Retrieved December 4, 2013, from http://c-spanvideo.org/program/FutureM

Smith, D. W. (2013). Phenomenology. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2013). Retrieved from http://plato.stanford.edu/archives/win2013/entries/phenomenology/

Snow, C. P. (1964). *The two Cultures: and a Second look* (2 ed.). C.U.P. Retrieved from http://capitadiscovery.co.uk/sussex-ac/items/22469

Solomonoff, R. J. (1968). The search for artificial intelligence. *Electronics and Power*, *14*(1), 8. https://doi.org/10.1049/ep.1968.0004

Sophocles. (2009). Antigone. Retrieved from http://classics.mit.edu/Sophocles/antigone.html

Sun, R. (2008). *The Cambridge Handbook of Computational Psychology* (1 edition). Cambridge ; New York: Cambridge University Press.

TheEconomist. (2013, May 14). Difference Engine: The caring robot. *The Economist*. Retrieved from http://www.economist.com/blogs/babbage/2013/05/automation-elderly

Togelius, J., Lucas, S. M., & Nardi, R. D. (2007). Computational Intelligence in Racing Games. In N. Baba, P. L. C. Jain, & H. Handa (Eds.), *Advanced Intelligent Paradigms in Computer Games* (pp. 39–69). Springer Berlin Heidelberg.

https://doi.org/10.1007/978-3-540-72705-7_3

Turing, A. M. (1953). Digital computers applied to games. In B. V. Bowden (Ed.), *Faster than Thought : a Symposium on Digital Computing Machines* (pp. 286–310). Pitman.

Turkle, S. (1984). *The second self: computers and the human spirit*. London: Granada. Retrieved from http://capitadiscovery.co.uk/sussex-ac/items/1155840

Turkle, S. (1991, March 17). Dangerous Thoughts . . . And Machines With Big Ideas. *The New York Times*. Retrieved from http://www.nytimes.com/1991/03/17/books/dangerous-thoughts-and-machines-with-big-ideas.html

van der Zant, T., Kouw, M., & Schomaker, L. (2013). Generative Artificial Intelligence. In V. C. Müller (Ed.), *Philosophy and Theory of Artificial Intelligence* (pp. 107–120). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-31674-6_8

Watson, I. (1999). Case-based reasoning is a methodology not a technology. *Knowledge-Based Systems*, *12*(5–6), 303–308. https://doi.org/10.1016/S0950-7051(99)00020-9

Watson, J. B. (1913). Psychology as the behaviorist views it. *Psychological Review*, *20*(2), 158–177. https://doi.org/10.1037/h0074428

Watson, J. B. (1914). *Behavior : an introduction to comparative psychology*. New York: H. Holt. Retrieved from http://archive.org/details/behaviorintroduc00watsuoft

Watson, J. B. (1920). Is Thinking Merely Action of Language Mechanisms1? (v.). *British Journal of Psychology. General Section*, *11*(1), 87–104. https://doi.org/10.1111/j.2044-8295.1920.tb00010.x

Watson, J. B. (1931). *Behaviorism* (Rev. ed.). London: Kegan Paul. Retrieved from http://capitadiscovery.co.uk/sussex-ac/items/24582

Watson, P. (2001). *Terrible Beauty: A Cultural History of the Twentieth Century: The People and Ideas that Shaped the Modern Mind: A History*. Phoenix.

Watson, P. (2006). *Ideas: A History of Thought and Invention, from Fire to Freud*. Harper Perennial.

Weizenbaum, J. (1966). ELIZA - a Computer Program for the Study of Natural Language Communication Between Man and Machine. *Commun. ACM*, *9*(1), 36–45. https://doi.org/10.1145/365153.365168

Wheeler, M. (2005). *Reconstructing the cognitive world: the next step*. London: MIT. Retrieved from http://prism.talis.com/sussex-ac/items/911756

Whitby, B. (2011). Do You Want a Robot Lover? The Ethics of Caring Technologies. *Robot Ethics: The Ethical and Social Implications of Robotics*, 233.

Winograd, T. (1971). *Procedures as a Representation for Data in a Computer Program for Understanding Natural Language*.

Winograd, T. (1991). Thinking Machines: Can There Be? Are We? In J. J. Sheehan & M. Sosna (Eds.), *The Boundaries of Humanity: Humans, Animals, Machines*. University of California Press.

Winograd, T., & Flores, F. (1986). *Understanding computers and cognition: a new foundation for design*. Norwood, N.J: Ablex. Retrieved from http://prism.talis.com/sussex-ac/items/272586

Wittgenstein, L. (2001a). *Philosophical Investigations: The German Text with a Revised English Translation: German Text, with a Revised English Translation* (3rd Edition edition). Malden, MA: Wiley-Blackwell.

Wittgenstein, L. (2001b). *Tractatus Logico-Philosophicus* (2 edition). London : New York: Routledge.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, *8*(3), 338–353. https://doi.org/10.1016/S0019-9958(65)90241-X

Zimmer, H. R. (1951). *Philosophies of India*. (J. Campbell, Ed.). Princeton, N.J.: Princeton University Press.